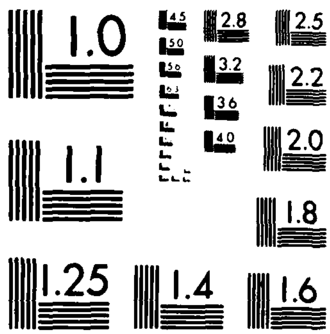
AD-A152 959 FORECAST VERIFICATION OF THE 107 CENTIMETER SOLAR FLUX 1/2
AND THE AP DAILY G..(U) AIR FORCE INST OF TECH
WRIGHT-PATTERSON AFB OH SCHOOL OF ENGI.. P M NOSTRAND
UNCLASSIFIED DEC 84 AFIT/GSO/PH-OS/84D-2 F/G 4/1 NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963 A

FORECAST VERIFICATION OF THE
10.7 CENTIMETER SOLAR FLUX AND THE
Ap DAILY GEOMAGNETIC ACTIVITY INDICES

THESIS

Philip M. Nostrand
First Lieutenant, USAF

AFIT/GSO/PH-OS/84D-2

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY
AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

FORECAST VERIFICATION OF THE
10.7 CENTIMETER SOLAR FLUX AND THE
Ap DAILY GEOMAGNETIC ACTIVITY INDICES

THESIS

Philip M. Nostrand
First Lieutenant, USAF

AFIT/GSO/PH-OS/84D-2

A-1

AFIT/GSO/PH-OS/84D-2

FORECAST VERIFICATION OF THE 10.7 CENTIMETER SOLAR FLUX AND THE Ap DAILY GEOMAGNETIC ACTIVITY INDICES

THESIS

Presented to the Faculty of the School of Engineering
of the Air Force Institute of Technology
Air University
In Partial Fulfillment of the
Requirements for the Degree of
Master of Science in Space Operations

Philip M. Nostrand, B.S.
First Lieutenant, USAF

December 1984

Preface

The impetus for this thesis has been the Space Command statement of work requesting a study of this particular problem. The cliche of being in the right place at the right time certainly applies to me since I was TDY to Colorado Springs the month the statement was released. I am grateful to Lt Col George Davenport, Chief of Aerospace Sciences at HQ 4WW for recognizing this potential thesis topic and for providing me with the names and phone numbers of many other people within Air Weather Service familiar with space environmental forecasting. These people in turn referred me to others, all of whom went out of their way to answer my questions, offer advice and mail me hard to find data, reports and papers. To thank all these people individually would require another chapter in this thesis. Instead I want to offer a collective thank you, so thank you everyone.

Although the power of the telephone proved invaluable in connecting me with some very accomodating people, this thesis certainly would have been incomplete without the assistance and advice of my advisors, Maj James Lange and Lt Col Joseph Coleman. Additionally, the support and encouragement of my friends and neighbors helped me keep my perspective and sense of humor these last few months, particularly Rob Hardy for teaching me how to spell persistence.

One final note; I wish to take personal responsibility for any misinterpretation of information which I referenced as a personal communication.

Table of Contents

List of Figures

List of Tables

Abstract

Air Force Global Weather Central space environmental forecasts of the 10.7 centimeter solar radioflux index and the Ap daily average geomagnetic activity index were compared with persistence "forecasts" to check for accuracy and skill. One, two and three day forecasts were compared. The AFGWC forecasts were found to be more accurate and skillful than the persistence forecasts.

The data base covered the period from 4 January 1971 through 29 April 1984. Statistics were calculated for the total data set and each individual year. Root mean square error and percentage of significant errors were used as measures of accuracy. A paired sign test was used to compare for skill. The test was run on significant errors and absolute errors. A significant error is when the difference between the forecast value and the verifying observed value (ie. the observation one, two or three days hence) is greater than ten.

The total data base yielded results which favored the AFGWC forecasts in all instances except one. The exception was the one day Ap sign test on absolute errors. Persistance also tended to do as well or better than the AFGWC forecasts on some individual years, primarily during the years around solar minimum (1975-1977) and also for the one day forecasts. It was found that AFGWC performed better at predicting a sudden decrease in the index values than it did predicting a sudden increase.

FORECAST VERIFICATION OF THE 10.7 CENTIMETER SOLAR FLUX AND THE Ap DAILY GEOMAGNETIC ACTIVITY INDICES

I. Introduction

Background

The upper atmosphere of the earth is a region dominated by complex interactions between the sun, the solar wind and the earth's geomagnetic field. Solar activity, indicated by disturbances in the geomagnetic field and increased solar radiation in radio wavelengths, causes heating of the upper layers of the atmosphere which in turn increases the number density of neutral particles in these layers. There are two mechanisms which contribute to the heating process. The first is direct heating by increased solar radiation, primarily by the extreme ultraviolet (EUV) and soft X-ray wavelengths. These wavelengths are hereby defined to fall between 100 and 1000 angstroms (Prochaska, 1984:3). The second mechanism is energetic particle heating. These charged particles (ions and electrons) are emitted from the sun and can take anywhere from three hours to three days to affect the upper atmosphere, while EUV heating occurs almost instantaneously.

There are many satellites in low earth orbits at altitudes of 200 to 700 km. During periods of solar activity, the atmospheric drag on these satellites increases because of the higher density. This process slows a satellite's vel-

ocity so that its orbit decays to lower altitudes. The predicted position of the satellites changes, causing tracking problems. As a result, spacetrack radar may detect an object where none is expected to be (Prochaska, et al, 1982:172).

The North American Aerospace Defense Command (NORAD) has a mission to know the position of all earth orbiting objects (Dept of the Air Force, 1984:1). Satellite tracking is also accomplished at the Air Force Satellite Control Facility (AFSCF) located at Sunnyvale Air Force Station. These missions entail monitoring the positions of satellites in low earth orbits and predicting the future positions of these objects. These predictions are also used to determine launch times and orbits for future spacecraft like the Space Shuttle.

NORAD uses atmospheric neutral density models which have been developed to aid in the orbit prediction of satellites. Density is a key parameter in a drag equation which calculates the deceleration of spacecraft (Prochaska et al, 1982:169). There are two important parameters in the density models: the solar flux and the geomagnetic activity index. Specifically, F10.7, the solar flux at the radio wavelength of 10.7 centimeters is used to indicate of EUV heating, and the Ap daily planetary average of geomagnetic activity is used to indicate heating due to injection of charged particles from the sun. Numerical values are required for both present and future times (Dept of the Air Force, 1984:1). How are the forecasts made and, more importantly, how accurate are they? This thesis will answer these questions.

Problem Statement

The Space Environmental Support System (SESS) branch of the Air Force Global Weather Central (AFGWC) has been forecasting solar flux and geomagnetic activity indices for over fifteen years. These indices are used in atmospheric density models by NORAD and the AFSCF to predict drag on low earth orbiting satellites. The problem is NORAD is unsure of the accuracy of the predictions. NORAD has been using persistence values as forecast parameters in their density models instead of using the SESS forecasts. Persistence is defined as predicting a continuation of the current situation. In other words, persistence uses today's observed values as tomorrow's forecast values.

Is there a statistically significant difference between forecast values versus persistence values for the F10.7 cm solar flux and the Ap daily planetary amplitude of geomagnetic activity? Do the forecast values or the persistence values come closer to the observed values for each forecast period?

Research Objectives

This research project will compare SESS forecast values and persistence "forecast" values to observed values for one, two and three day forecasts. The primary objective of this thesis will be to determine whether NORAD should use SESS Ap and F10.7 forecasts.

Specifically, I shall test the null hypothesis that there is no difference between the accuracy of SESS fore-

Table 2-4

Magnetic Observatories Used by AFGWC

Observatory	Geographic Lat	Geographic Long	Geomagnetic Lat	Geomagnetic Long
Boulder, Colorado	40 08N	105 14W	+49.0	316.5E
College Observatory, Fairbanks, Alaska	64 52N	147 50W	+64.6	256.5E
Goose Bay, Labrador, Canada	55 20N	60 30W	+60.5	11.9E
Loring AFB, Maine	46 57N	67 53W	+58.5	1.5E
RAF Upper Heyford,	51 56N	1 15W	+50.7	79.1E

(Prochaska, 1980:6)

Table 2-5

Magnetic Observatories Used to Determine Gottingen Ap

Lerwick, Shetland Islands	60 08N	358 49E	+62.5	88.6E
Lovo, Sweden	59 21N	17 50E	+58.1	105.8E
Sitka, Alaska	57 04N	224 40E	+60.0	275.4E
Rude Skov, Denmark	55 51N	12 27E	+55.8	98.5E
Eskdalemuir, Scotland	55 19N	356 48E	+58.5	82.9E
Meanook, Canada	54 37N	246 40E	+61.8	301.0E
Wingst, West Germany	53 45N	9 04E	+54.5	94.0E
Witteveen, Netherlands	52 49N	6 40E	+54.2	91.0E
Hartland, England	51 00N	355 31E	+54.6	79.0E
Agincourt, Canada	43 47N	280 44E	+55.0	347.0E
Fredericksburg, Virginia	38 12N	282 38E	+49.6	349.9E
Amberly, New Zealand	43 09S	172 43E	-47.7	252.5E

(Prochaska, 1980:13)

Ap from three-hourly K values. This report is believed to be the first step in the creation of the current five station network of observatories which contributes to the daily calculation of an Ap value. The AF Ap is computed at the Air Force Global Weather Center (AFGWC) at Offutt AFB, Nebraska. Prochaska (1980) has written a valuable guide which describes the calculation of a number of magnetic indices at AFGWC, including Ap. All further references to Ap will be to the AF (real-time) Ap, as distinguished from the Gottingen Ap.

There are some important differences between the two Ap indices. The observatory networks are different. Table 2-4 lists the location of AFGWC observatories and Table 2-5 lists the location of observatories used to determine Gottingen Ap. Most of the AF stations are in North America (4 out of 5), while ISGI stations are mainly in Europe (7 out of 13). Different observatories in different geographic distributions present complications both within and between each network. Diurnal, seasonal, and latitudinal effects must be accounted for since the amount of a geomagnetic disturbance varies with respect to all three of these factors. Each Ap attempts to standardize separate station readings before averaging but their methods are different (Prochaska, 1980:8-10; Allen and Feynman, 1979:391-392). Furthermore, the standardization of diurnal variations by ISGI has been criticized (Allen and Feynman, 1979:392).

It is somewhat surprising that very few published studies have been done to compare the two Ap indices. Coommuni-

Table 2-3

Relationship Between the Values of Kp and ap

Kp	ap	Kp	ap	Kp	ap	Kp	ap
0o	0	2+	9	5-	39	7o	132
0+	2	3-	12	5o	48	7+	154
1-	3	3o	15	5+	56	8-	179
1o	4	3+	18	6-	67	8o	207
1+	5	4-	22	6o	80	8+	236
2-	6	4o	27	6+	94	9-	300
2o	7	4+	32	7-	111	9o	400

(Rostoker, 1972:940)

are not used very often (Rostoker, 1972:936-940).

There are actually two versions of Ap in existance. The most widely accepted value of Ap is produced by the International Services of Geomagnetic Indices (ISGI) in Gottengen, West Germany. These values are derived from a worldwide network of 13 magnetometer stations (Allen and Feynman, 1797:388). The Gottengen Ap values are not available to the scientific community in real-time; there is a lag of at least one month before they are released (Dandekar, 1982:8; Schleher, 1984).

In the early 1960's, a USAF/AWS report was released identifying a requirement "for the availability on a daily basis of the planetary Ap index of geomagnetic activity" (Dept of the Air Force, 1963:1). This report applied regression analysis techniques to select the best combination of North American magnetic observatories for use in estimating

of authors have commented on this attribute (Chernosky, 1965; Fraser-Smith, 1972; Prochaska, 1980). Ap has units of 2 gammas where 1 gamma = 10^{-5} Gauss = 10^{-9} Tesla (Prochaska, et al, 1981:137). Values of Ap can range from 0 to 400 although values above 100 are very rare.

Ap may be regarded as the 24 hour average of eight three hourly ap indices. The ap index is an average of individual magnetometer station measurements, ak. This ak is a measure of one-half the amplitude of the largest fluctuation of the earth's magnetic field at a single location for that particular three hour period (Rostoker, 1972:938).

It is important to explain the various A- and K-indices to avoid confusion in the next few pages. The three hourly station amplitude, ak, is directly related to the K index. The K index is also a three hourly index of activity but the maximum amplitude is converted to a quasi-logarithmic scale which is then standardized to account for the latitude of the station, the lccal time of day and the season. The standardized K values (Ks) are then averaged for all stations to obtain the Kp value. Kp is therefore related to the ap, the three hourly planetary amplitude. The values of Kp range from 0 to 9 and are broken into thirds (-, o, +). There are 28 values which the Kp may have: 0o, 0+, 1-, 1+, ..., 8+, 9-, 9o. The numerical relationship between Kp and ap values is presented in Table 2-3. There is no complementary K index for the Ap index. Kp has been summed over the eight daily three hourly periods but these values are not as easily interpreted as the Ap and

Geomagnetic Indices

> In general terms, an index is a quantity which provides a means of summarizing an otherwise detailed set of observations which are required to thoroughly describe a given process.
> (Gorney and Mizera, 1983:1).

A number of indices have been developed to measure geomagnetic activity, and the literature is rich with descriptions of these indices (Chernosky, 1965; Rostoker, 1972; Allen, 1982; Allen and Feynman, 1979). Each index has a particular use or else is related to a particular geographic region. Some indices reflect conditions in the auroral zone (the AE index) while others measure mid-latitude geomagnetic activity (K and A indices). In general, the number and complexity of indices have increased as scientists' understanding of the workings of the magnetosphere and the interaction between the magnetosphere and the interplanetary medium has increased.

One of the most used geomagnetic indices is the daily equivalent planetary amplitude index, hereafter referred to as the Ap (read as A-sub-p). According to Rostoker, in a review purposely written to "define the origin and status of a number of frequently used indices," the Ap was first introduced by Bartels in 1951 (Rostoker, 1972:936,938). Ap was created as a linear-scaled sister index to the older Kp index which is a quasi-logarithmic number used to characterize the level of worldwide geomagnetic activity at sub-auroral latitudes.

The linearity of Ap makes it more useful than Kp in mathematical calculations and correlation studies. A number

Readers interested in learning more about the international state of space environmental forecasting are referred to the Solar-Terrestrial Prediction Proceedings (Donnelly, 1979b), a four volume set of presentations made at the Solar-Terrestrial Prediction (STP) Workshop held April, 1979, at Boulder, Colorado. This work contains numerous papers describing forecast methods at space environment support centers around the world, plus user requirements, current and future needs, etc. One of the goals of the workshop was to "provide indepth interaction of prediction users, forecasters and scientists involved in the research and development of prediction techniques" (Donnelly, 1979a:v). In this respect, the workshop represented a large scale opportunity to examine user needs and how forecasters can better satisfy those needs.

In March, 1982, a similar workshop was sponsered by the SESC to specifically address the needs of users who are "adversely affected by solar-induced fluctuations in the neutral atmospheric density at high altitudes" (Joselyn, 1982b:v). The proceedings from these two workshops provided the bulk of literature reviewed for this thesis. A second STP Workshop was held in France during June, 1984. Although the proceedings have not been published at this time (Nov 1984), the author did receive a few of the presentations from one of the participants. These papers focus mainly on ionospheric forecasting and are not directly applicable to this report. The proceedings may prove useful for follow-on research, however.

levels and solar radiation levels are both point forecasts and will be discussed in depth in the next sections of this chapter. It is appropriate to note here that the A and K indices are subjective forecasts while the ten-centimeter flux is a subjective interpretation of a regression algorithm (Heckman,1979:330-341).

Table 2-2

Prediction Products (Lead Time Given in Parentheses)

<u>LONG TERM SOLAR ACTIVITY AND SOLAR RADIATION LEVELS</u>

- Smoothed sunspot number (1 month-10 years)
- Geomagnetic activity and ten-centimeter flux (1 month-10 years)
- General level of solar activity (27 days)

<u>SOLAR ACTIVITY -- SHORT TERM</u>

- Solar Flares (1, 2, 3 days)
- Solar proton events (1, 2, 3 days, PFP*)

<u>SOLAR RADIATION LEVELS -- SHORT TERM</u>

- Ten-centimeter flux (1, 2, 3 days)

<u>GEOMAGNETIC DISTURBANCE LEVELS</u>

- A, K-indices (1, 2, 3 days)
- Time of sudden commencements (PFP)
- Storm size (PFP)

---

* PFP: Post Flare Prediction - A prediction of a flare consequence once the flare has occurred.

(Heckman, 1979:330)

Table 2-1

SESC Users and Types of Activity Affecting Their Systems

Customer	Type of Activity Producing Effect
- Civilian satellite communication	- Magnetic storms
- Commercial aviation--mid-latitude communication (VHF)	- Solar radio emissions
- Commercial aviation--polar cap communication (HF)	- PCA, magnetic storms, x-ray bursts
- Commercial aviation navigation (VLF)	- PCA, magnetic storms, x-ray bursts
- Electric power companies	- Magnetic storms
- Long line telephone communication	- Magnetic storms
- High altitude polar flights--radiation hazards	- Solar proton events
- Civilian HF communication Coast Guard, commercial companies, GSA, VOA	- X-ray emission, UV emission magnetic storms
- Geophysical exploration	- Magnetic storms
- Satellite orbital variation	- UV emission, magnetic storms
- DOD SATCOM communication	- Magnetic storms
- DOD HF communication	- X-ray emission, UV emission PCA, magnetic storms
- DOD reconnaissance	- PCA, magnetic storms
- DOD navigation	- X-ray emission, UV emission
- ERDA communication (prospective customers)	- X-ray emission, UV emission magnetic storms
- International community	- All
- Scientific satellite studies: IMS, Solar Maximum mission, Shuttle, solar constant measurements, stratospheric ozone variation, interplanetary missions	- Optical solar flares, magnetic storms, x-ray emission, UV emission, solar proton events, solar features
- Scientific rocket studies IMS, magnetosphere, ionosphere, upper atmosphere, sun	- Optical solar flares, solar features, magnetic storms, solar proton emission, x-ray emission
- Scientific ground studies IMS, sun, interplanetary magnetosphere, ionosphere, upper atmosphere, stratosphere, troposphere, seismological/geomagnetic	- Optical solar flares, magnetic storms, solar proton emission, x-ray emission, UV emission, solar features

(Heckman, 1979:323)

razor, also known as the principle of parsimony." Indeed, that is the issue addressed by this work: is persistence, a simple forecast method, equally as good as the forecaster's predictions?

In the field of solar-terrestrial predictions, there are many customers affected by a variety of solar and/or geophysical disturbances. The customers are supported by a much smaller network of forecast centers. While it is beyond the scope of this research to describe all users of all forecasts produced, Table 2-1 is included to give an example of the number of customers and the types of activity affecting their systems supported by the Space Environment Services Center (SESC) in Boulder, Colorado. It is interesting to note, with respect to this thesis, that practically all users are affected by events indicated by magnetic storms, and almost half are affected by solar radio or ultraviolet emissions.

Table 2-2 lists the prediction products and lead times routinely provided by the SESC. These products cover the gamut of forecast types. The long-term forecasts use quantitative techniques: statistical analysis for sunspot numbers and regression equations for geomagnetic activity and ten centimeter predictions. The 27 day general level of activity are categorical forecasts: five levels for solar activity (very low to very high) and five levels for geomagnetic activity (quiet to major storm). Solar flares are forecast as the probability of occurrence for each of the three separate classes of flares. Geomagnetic disturbance

into one of at least two mutually exclusive categories. Rain/no rain and intervals of cloud heights are both categorical forecasts. Finally, predictions may be made on the probability of occurrence of a future event or condition. The percent chance of rain is a frequently used probability forecast (Brier and Allen, 1951:843-846).

Since forecasts are generally not made by the people who use them, there needs to be a certain amount of interaction between the forecasters and their "customers." The needs of the user must be established. What does the user want to know, how often does he or she need the prediction, what is the forecast horizon, what degree of accuracy is required or acceptable? These are all questions to answer before forecasting can start (Abraham and Ledolter, 1983:4). The forecast horizon is the time or period for which the forecast is made, for example tomorrow or next week.

Sometimes what can be forecast is not what is needed. Other times what is needed cannot be forecast. The availability of data may influence the type and accuracy of a forecast which can be made. A point forecast may be preferable to the user versus a categorical forecast, but the cost of collecting data, developing models and producing the preferred type may outweigh its usefulness. Since the basic objective of forecasting is producing forecasts which are seldom incorrect, accuracy is the most important attribute for choosing a particular method. Given two equally able forecast methods, however, the simplest one is preferred. Abraham and Ledolter (1983:5) refer to this as "Ockham's

predictions determine rates of production and possibly whether to hire or fire employees. Technological predictions influence the time to modernize assembly lines, automate office operations or acquire military equipment. This thesis is about forecasts of space environmental conditions about which some readers may not be intimately familiar. As an aid to understanding, forecast and verification discussions will occasionally use examples away from the specific field of the space environment.

There are two broad types of forecasts: qualitative and quantitative. A qualitative forecast is a person's subjective interpretation of pertinent data to arrive at an intuitive prediction. Two forecasters analyzing the same situation will not necessarily come up with identical forecasts. A quantitative forecast on the other hand is an objective prediction, generally using statistical or mathematical methods to analyze the data and yield a prediction. This type of forecast is frequently made using automated techniques. Given the same starting conditions, a quantitative method will produce the same forecast every time. Regression equations and time series models are common quantitative forecasts (Abraham and Ledolter, 1983:2-3).

Within each of these types, a prediction may be further broken down into three classes: point, category or probability forecasts. A point forecast is the prediction of a specific number. Tomorrow's high temperature and next month's unemployment rate are examples of point or numerical forecasts. A categorical forecast places the future event

II. LITERATURE REVIEW

This chapter will be broken down into four broad sections. The first will introduce the subject of forecasting in general and solar-terrestrial forecasting in particular. The next two sections will discuss the geomagnetic and solar flux indices respectively. They will include discussions of physical meaning, use, forecast history and advantages/disadvantages. The final section will be about verification: its purposes, the process of evaluation, control forecasts and the difference between accuracy and skill.

Forecasting

A forecast can be defined as a prediction of a future event or state. The prediction is usually made based on some type of analysis of relevant data. The objective of forecasting is to reduce forecast error (Abraham and Ledolter, 1983:1,5). Forecast error is the difference between what was forecast and what was subsequently observed at the time or period the forecast was made for.

Forecasting is frequently associated with weather. This is not unusual since most people depend on tomorrow's expected weather conditions to aid in the decisions of what clothes to wear, when to leave for work, etc. However, weather is not the only thing that is forecast. Many businesses and governmental organizations depend on numerous predictions for a variety of purposes. Economic forecasts are used to plan budgets and set interest rates. Sales

4. It will be assumed that changes in the network of magnetometer observatories used to calculate the Ap values will not impact the quality of the data.

5. It will be assumed that the average skill level of the SESS forecasters has remained constant during the period the data base encompasses.

Preview

Chapter II contains a literature review focusing on the indices of interest, their definition, their behavior with respect to the solar cycle, their uses, their forecasting history and their drawbacks and disadvantages. Suggestions for improvement and alternate indices also are made. The literature review will also highlight the field of forecast verification, including reasons for verification, problems with verification and persistence as a baseline to compare forecasts.

Chapter III is a discussion of the methodology including assumptions made about the data and criteria used for comparing the forecasts (ie. the specific hypotheses). Results and analysis are covered in Chapter IV. Conclusions and recommendations are presented in Chapter V.

casts and persistence. The alternate hypothesis is SESS forecasts are more accurate than persistence. Additionally, a test will be conducted to determine whether SESS forecasters exhibit a significant amount of skill in their predictions when compared to the unskilled method of persistence.

The testing will be conducted over a data base covering the last eleven year solar cycle. This will include tests over the whole period and tests over each year (or portion of a year) of the data. The annual testing will be done to determine if there are variations in accuracy during solar maximum, solar minimum, increasing solar activity and decreasing solar activity.

Scope and Limitations

1. The data base will consist of observed values and one, two and three day forecasts of Ap and F10.7 indices. Data is from 1 January 1971 to 29 April 1984. All the observed values and the forecasts from 1975 were obtained from a data tape sent from the Air Weather Service (AWS) detachment at Sunnyvale AFS. The early forecast values were sent in hard copy from the Space Environmental Service Center at Boulder, Colorado.
2. This data covers the declining phase of solar cycle 20 to the declining phase of solar cycle 21.
3. The Ap observed values used to verify the forecasts will be the Air Force produced real-time numbers rather than the Gottengen values which are not published in real-time.

cation with AWS personnel reveal an unpublished study which had been done that could not be located. However, it is believed the results indicated that the difference between the two is "within the noise level" (Dye, 1984). Dandekar (1982:11) did examine the correlation between the AF and Gottengen Kp indices using data from March 1978 to May 1981. The correlation coefficient was 0.837, which indicates a "good" although far from perfect agreement between the indices. Heckman (1979:328) writes that AF Ap values "are frequently made artificially large by the effects of auroral electrojets on their station sample" although he does not cite the study which makes that conclusion. Patterson (1984) raises the issue that, in the long run, trying to determine how close AF Ap is to Gottingen Ap is really a moot question. Both indices are deficient, since neither has a network of ideally placed observatories (ideal being defined as uniform longitudinal spacing at one geomagnetic latitude).

The behavior of Ap with respect to the solar cycle has been investigated. There is general agreement among investigations. Fraser-Smith examined the periodic variations of geomagnetic activity and of sunspot numbers over a 38 year period (1932-1970). He conducted a spectral analysis using monthly averages of Ap and monthly averages of sunspot numbers. Conclusions of this study include that Ap fluctuations are much "noisier" than sunspot numbers and that large maximums of Ap occur during the declining phase of the solar cycle. These features had been observed before and were

believed to be associated with "M-region" activity on the sun. Additionally, a spectral line for Ap at 27 days was observed, which is the sun's rotational period, but no corresponding sunspot line was found (Fraser-Smith, 1972:4211,4218).

Since 1972, increased observations and understanding of the sun has identified M-regions as coronal holes. Heckman (1979:340-341) discusses recurrent activity of both coronal holes and associated geomagnetic disturbances with respect to forecasting. It is generally acknowledged that coronal holes are most long lasting during the years between solar maximum and solar minimum, and the behavior of magnetic indices is dominated by the recurrent coronal hole disturbances (Prochaska, et al, 1981:140; Allen, 1982:114-118). During solar maximum, geomagnetic fluctuations are larger because of solar flares, but there is generally not quite as many large fluctuations, or major storms. Figure 1 shows the number of days with very high Ap compared to sunspot numbers. Note that the year with the most large magnetic storms during a sunspot cycle is after the year of sunspot maximum.

There is one additional periodic geomagnetic phenomena; the semi-annual variation. According to Allen (1982:116), this is not associated with the solar cycle, but rather the two annual periods when the magnetosphere is best orientated for coupling with the interplanetary magnetic field (see figure 2). This is a good example of how an index can prove to be an indicator and aid in science's understanding of how the complex sun, solar wind, earth system interacts.

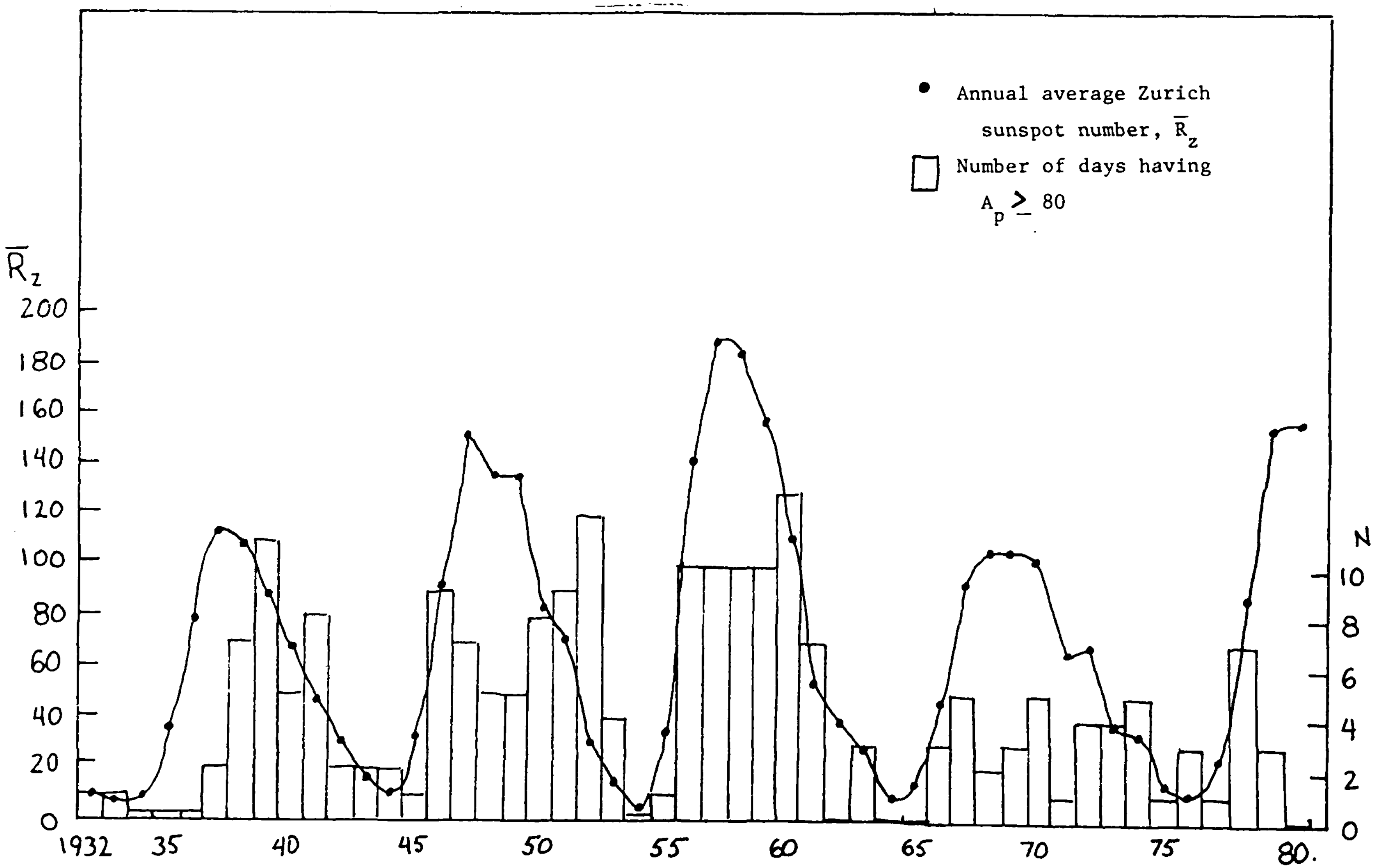


Figure 1. Annual Number of Large Magnetic Storms - count of large (Ap $\geq$ 80) magnetic storms in each year compared with smoothed annual sunspot number from 1932-1980. (Allen, 1982:117)

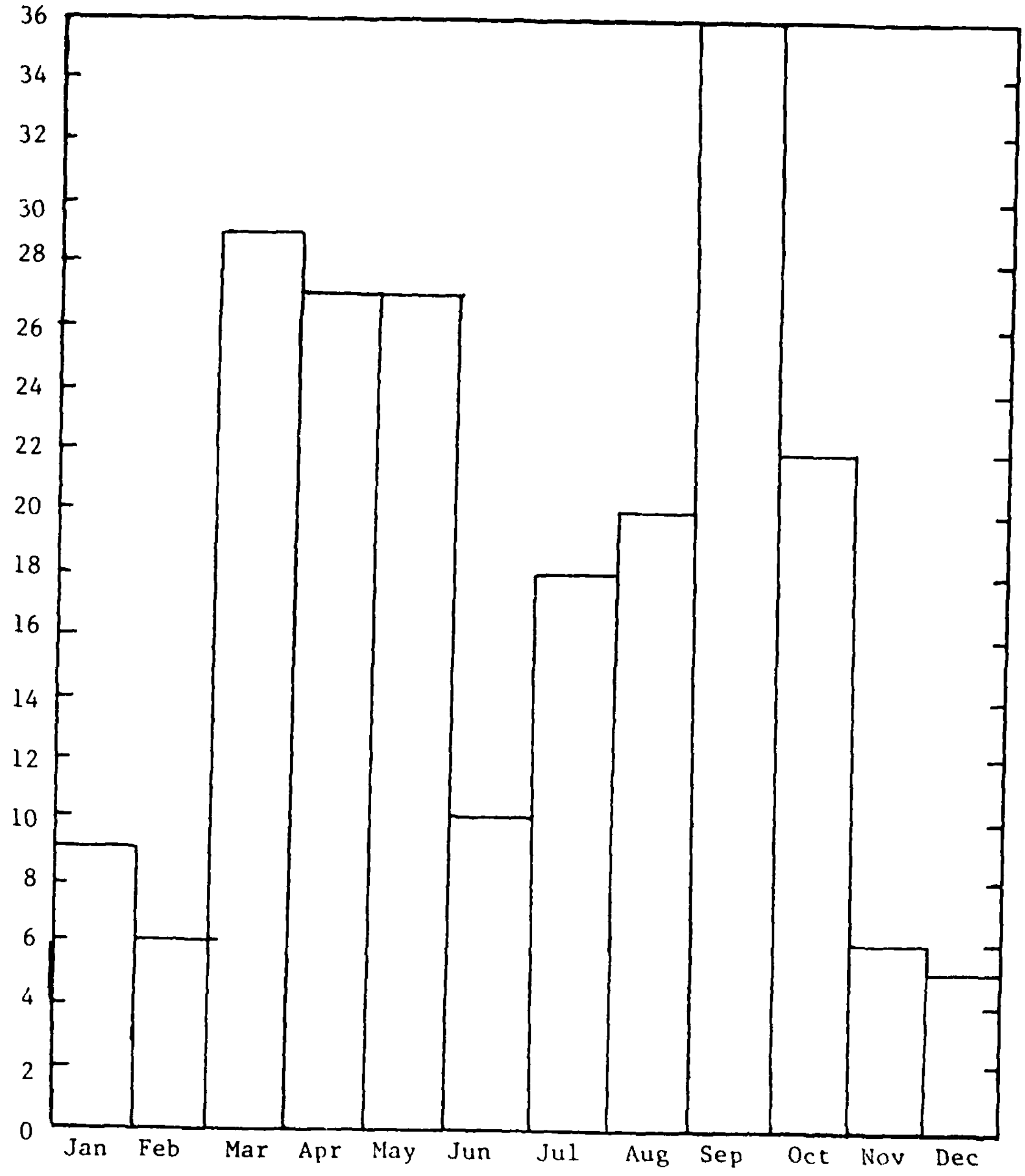


Figure 2. Seasonal variation in cumulative number of large ($Ap \geq 80$) magnetic storms, 1932-1980.

(Allen, 1982:118)

Since geomagnetic activity has numerous effects on man and man-made systems, the various indices have been used to relate the effect to the systems. This section discusses the uses of Ap.

The most relevant use of the Ap index with respect to this thesis is as an indicator of joule heating by charged particles in the development of atmospheric density models. Works by Lean (1982) and Vampola, et al (1979) discuss Ap input into neutral density models and the need for accurate predictions of Ap when the models are used to predict satellite orbit evolution. These papers represent summaries from workshops held to address the topics of satellite drag and solar-terrestrial predictions respectively. Their bibliographies and the other papers they summarize are rich in content of how Ap is used and how upper atmospheric heating could perhaps be better represented by other indices.

Atmospheric density modeling is not the only phenomena related to geomagnetic activity. Electric power transmission and oil and gas pipelines are affected by geomagnetically induced ground currents (Campbell, et al, 1979:133-135). Warnings of expected periods of strong activity can alert operators of potential system failures. Geological mapping sensors need periods of quiet activity to aid in the searching for ores and minerals, while radar and communications systems must be adjusted during magnetic storms (Paulikas and Lanzerotti, 1982: 42-46). Forecasts of such events are required.

The last few paragraphs have highlighted the need for

geomagnetic forecasts. This does not imply that there is no current forecasting ability. Indeed, Ap has been forecast by the USAF since the late 1960's (Thompson and Secan, 1979:351). The following is a discussion of who makes the forecasts, what techniques are applied to the forecasts, limiting factors, and what may be done to improve the forecasts.

The Space Environment Services Center (SESC), located in Boulder, Colorado is the source of Ap forecasts in the United States. The SESC is jointly operated by the National Oceanic and Atmospheric Administration (NOAA) and the Air Force AWS. Forecasts are issued jointly; however, AFGWC is responsible for the Ap forecasts issued daily for the next one, two, and three days. Weekly seven day forecasts and 31 day forecasts are also made by AFGWC. (This information was compiled from papers by Joselyn (1982a), Heckman (1979), and Thompson and Secan (1979) and comfirmed by personal discussion with Ashton (1984)).

The forecasts are made for the AF Ap value, not the Gottengen Ap. They are point forecasts for specific Ap values and are subjectively generated. In other words, there is no quantitative model used to produce the forecast values.

Observations of the sun and the near-earth solar wind environment are used to aid in the predictions. Joselyn (1982a:140-141) describes three types of solar observations useful in predicting geomagnetic activity: flares, coronal holes, and disappearing solar filaments (DSF). Unfortu-

nately, observations of these phenomena do not mean that a magnetic disturbance will definitely result.

Townsend (1984) has said that the scientific increase in understanding of the solar, solar wind and magnetospheric interaction which has occured during the last 10 to 15 years has not necessarily improved a forecaster's ability to predict geomagnetic indices. Prochaska, et al (1981:137-143), have written an extensive discription of subjective forecast techniques for recurrent (coronal hole and current sheet crossing) and flare induced disturbances.

Since the theory has improved, why do some believe "magnetospheric forecasting at AFGWC is still in its infancy" (Thompson and Secan, 1979:354)? Forecasters are limited by incomplete understanding of solar wind, magnetosphere, ionosphere interactions; unavailability of solar wind data; and lack of models which can use currently available space based data (Thompson and Secan, 1979:363-364; Joselyn, 1982a:142). The ISEE-3 satellite had been used by the SESC to measure the solar wind and interplanetary magnetic field; this aided in short term (30 minute) warnings of solar wind changes, and in identifying the direction of the interplanetary magnetic field (IMF) (Joselyn, 1982a:142). Unfortunately, the spacecraft was moved from its optimal solar wind viewing location within the past two years.

This section will conclude with a brief discussion of the disadvantages to the Ap index and to its "forecastability". Alternatives (better indices) will be suggested and the reasons why they are not now being used will be

highlighted.

The biggest drawback to Ap is that it is essentially outdated. Allen and Feynman (1979:391) state that since Ap was devised so long ago, "it was not designed to measure certain specific processes we now envision as basic magnetospheric dynamics, such as the enhancement of the ring current or substorms." It is their belief that its usefulness has been stretched about as far as possible. While it does have some ability to detect the type of IMF-magnetosphere interactions now believed to exist (see Figure 2), it cannot accurately explain phenomena which occur in smaller temporal and spatial scales. They advocate the use of alternate indices (AE) or the possibility of "directly monitoring interplanetary conditions in the solar wind and thereby making possible realtime predictive applications, possibly removing the need for some indices" (Allen and Feynman, 1979:385).

Another problem with Ap is the poorly distributed network of observatories. Rostoker (1972:944) remarks that the wide longitudinal gaps and mid-latitude alignment of the stations make Kp a poor "quantitative indicator of the intensity of a given substorm or level of substorm activity." He also remarks that Kp and Ap do have value when used in "statistical analyses of long periods of magnetospheric activity for the purpose of determining long-term trends."

Heckman (1979:328) has questioned the spatial and temporal resolution problems with A- and K- indices. The

global nature of these indices restricts their applicability, while the 90 minute minimum reporting interval is too large to be of use to the electric power industry which is concerned with large variations within time periods of a few minutes.

The AE (Auroral Electrojet) index has some advantages over Ap. While it is also a global index, its stations are better distributed at higher (auroral) latitudes and observations are recorded much more frequently (2.5 minutes vs 3 hours for ap). It is therefore able to indicate substorm activity better, although this ability is still not perfect (Rostoker, 1972:940-942,945-946). Additionally, although observations are frequently made, this index is not available in real-time due to data reduction problems. It suffers a publication delay similar to the Gottengen Ap (Allen, 1984).

The work by Akasofu has resulted in a parameter, epsilon, which determines the energy available for release as a geomagnetic (sub)storm, this energy previously being stored in the magnetosphere. Epsilon is a function of solar wind parameters, and in a recent paper he describes a numerical forecasting scheme for geomagnetic storms given the ability to predict epsilon as a function of time (Akasofu, 1984). This scheme could be used to predict AE and Dst (another magnetic index which indicates low latitude activity) (Akasofu, 1984:3). This work appears promising and AWS is monitoring its progress (Schleher, 1984).

Gorney and Mizera (1983) proposed the development and

use of a new index called the Total Auroral X-Ray Intensity (TAXI) index. It is calculated from data collected on the Defense Meteorological Satellite Program (DMSP) satellite. The data measures energy at the x-ray wavelength of electrons precipitating into the upper atmosphere (500 km). A study was conducted by personnel from Sunnyvale and the Air Force Geophysics Laboratory (AFGL) to compare it with Ap values. The TAXI index did not perform significantly better than Ap according to one of the participants (Townsend, 1984), although the data base was not very large (4 individual weeks in 1983).

The primary reason why Ap is still in use even though more specific indices may be available to the users is economic. Users, especially people who develop and run atmospheric density models, have large investments in their models. An alternate index would require that the models be reconstructed (Heckman, 1979:328). According to a summary report from the Satellite Drag Workshop, "...the absence of a long historical data base would cause difficulties for model fitting, and in addition, users would need assurance that the data would be continuously available in the future" (Richmond, 1982:97). Additionally, there is a cost factor involved in either producing a new real-time index or building a satellite to receive space based observations. Ap has the dual advantage of a large data base and a relatively low-cost existing ground based observation network.

Solar Flux Index

The 10.7 centimeter solar flux (hereafter refered to as the F10.7) has a much simpler description than the Ap geomagnetic index. Prochaska gives a concise description of several aspects of F10.7: its forecast history, including original and current regression equations; its "forecastability," ie. potential for improvement over the regression equations based on solar activity observations and theory; and its usefulness in predicting drag (Prochaska, 1984:3-5). This section will expand on the above list, using additional references as necessary.

The usefulness of F10.7 as an indicator of extreme ultraviolet (EUV) heating of the upper atmosphere has been well documented throughout the literature (Prochaska, et al, 1981; Vampola, et al, 1979; Heckman, 1979; Prochaska, 1984). It can be classified as an index since it does not measure heating per se, but only serves as an indicator of heating.

The F10.7 measurement involves observation of emission of radiowaves at the 10.7 centimeter wavelength (2700 MHz frequency) from the entire solar disk. EUV and 10.7 radiowaves originate from layers close to each other in the sun's chromosphere and have been shown to be statistically correlated (Prochaska, 1984:3). However, EUV energy is absorbed by the earth's atmosphere and cannot be measured at the ground, while F10.7 can be measured by ground-based instruments. Therefore, F10.7 is used as a "measure" of atmospheric heating (Prochaska, 1984:3).

The Canadian National Research Council has been measur-

ing F10.7 daily since 1947 with an antenna-type radiometer situated near Ottawa. The 1700Z (local noon) measurement has been accepted as the world standard and is used by many people involved with radio communications and upper atmospheric conditions. Heckman (1979:327), in describing forecasting at the SESC, claims that "this parameter is probably the most frequently requested and used parameter in the field of solar-geophysical measurements."

Flux, or irradiance, is the quantity determined by these measurements and is the rate of energy received per unit area per unit time. Solar radio flux is measured in Solar Flux Units (SFU), where 1 SFU = 1×10^{-22} Watts/meter^2. F10.7 indices are integer values of the 10.7 measurement results in SFU and range from less than 70 SFU to over 300 SFU (Prochaska, 1984:4).

Studies of F10.7 have shown that it has two periodic cycles of variation, both associated with the behavior of the sun. The "basic component" of variation is related to the 11 year solar cycle. Figure 3 is a graph of this phenomenon. Note here that solar minimum is not exactly halfway between the two maxima, but that the decreasing phase is longer (by two to three years) than the increasing phase. The "slowly varying component" has a period of approximately 27 days which corresponds to the period of one solar rotation. This recurrent activity is believed to be related to the reappearance of long-lived coronal holes (Prochaska, et al, 1981:72). Figure 4 is a graph of this phenomenon.

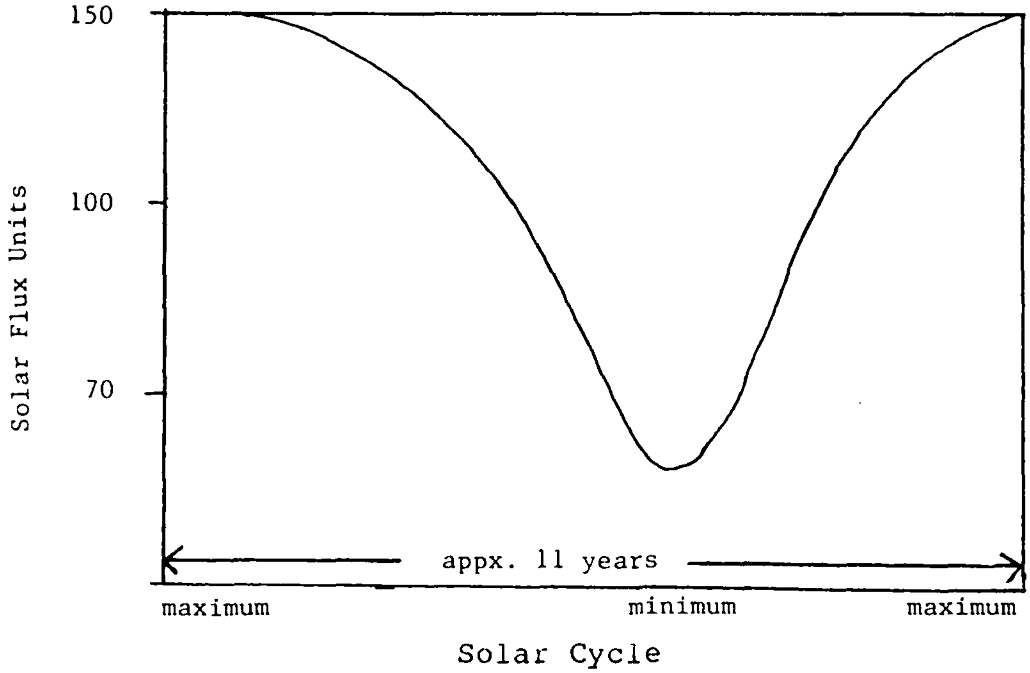


Figure 3. Quiet Sun Radio Variations

(Prochaska, et al, 1981:71)

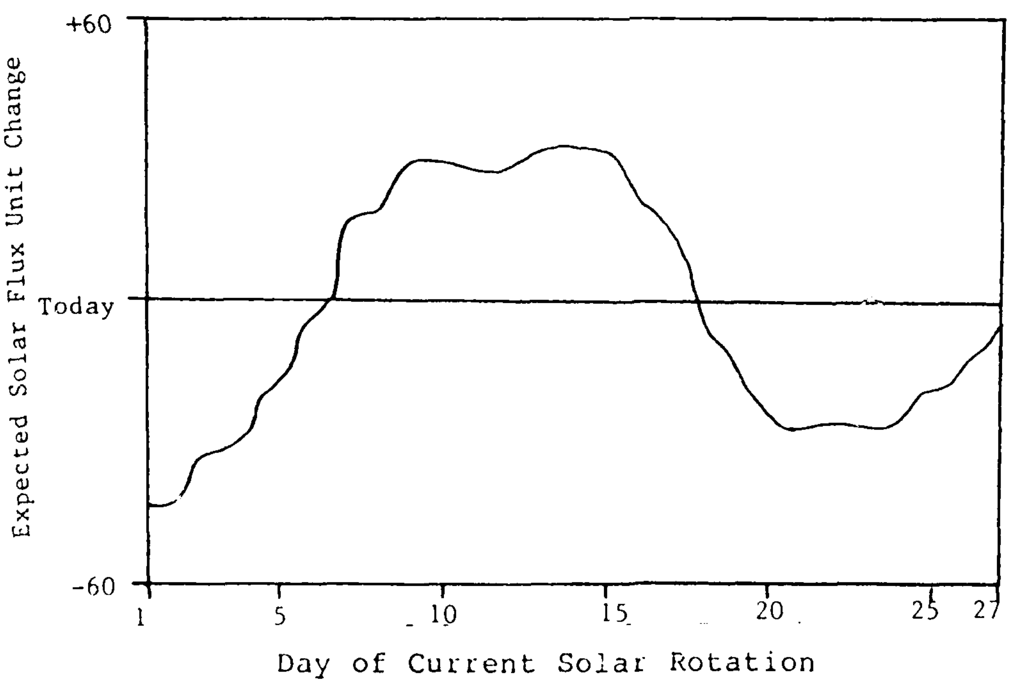


Figure 4. Slowly Varying Component

(Prochaska, et al, 1981:72)

When the F10.7 is averaged over a number of solar rotations, the resulting mean value is called the "disk component." When this component is used as an indicator of solar disk variations, it better predicts effects on the earth than the daily values which indicate active region variations (Vampola, et al, 1979:13). Every day AFGWC issues a previous 90-day mean F10.7 value as a measure of the total disk emission level (Prochaska, et al, 1981:169).

Daily variation of F10.7 is the result of evolving active regions on the sun. Prochaska, et al (1981:74-79), describe the different types of regions and the associated patterns of growth and decay.

Like the Ap, the F10.7 has many uses because solar EUV and x-rays affect many operations in space. The effect on the neutral density of the atmosphere is of importance to those concerned with drag and orbital lifetimes of low altitude satellites (Vampola, et al, 1979:13; Prochaska, 1984:4). However, UV and radio emissions also cause radio interference. Communication and radar systems are affected by increased noise, attenuation of signal and beam scattering (Prochaska, et al, 1981:75). Table 2-1 lists those SESC customers affected by solar a radio emissions, x-ray bursts, and x-ray and UV emissions. Forecasts are needed by these customers.

The forecasting of F10.7 in the United States is also a simpler process than predicting Ap. This is because the predictions, while still subjective, are made based on a series of regression equations run daily by AFGWC. These

sent for the years 1971 through 1983 while F10.7 data was sent for the years 1971 through 1975 only. Beginning in 1980, the Ap data pages included the monthly verification statistics.

It is of interest to describe the makings of the Sunnyvale data tape. Each day, the Operations Branch of Det 3 receives various reports from the SESC as messages transmitted via the CONUS Meteorological Data System (COMEDS). These reports include: eight three hourly ap and resultant Ap values; one F10.7 observed value; and one report with Ap and F10.7 1, 2 and 3 day forecast values. This data is recorded in a log book and also on AF Form 1530, Punch Card Transcript. Every four or five months the transcripts are collected by the Environmental Simulations Branch, and the data is entered onto punch cards and then read onto magnetic tape.

Two types of error are possible from this data handling process. One type of error may result when changes are made in the ap of Ap values sent to Sunnyvale from Boulder. This occasionally happens when magnetometer data from one of the five AF observatories is received late. ap values are calculated and disseminated every three hours using currently available data. If an observatory is late (due to equipment problems, communication problems, etc.), ap is calculated without that observation. The ap and Ap values will then be recalculated and retransmitted when the observation is reported. Occasionally, the timing of a late observation is such that the Ap value recorded in the log is affected. The

III. Methodology

The methodology chapter is divided into three sections. The first section will discuss the data base. Section two will review the methods of determining the statistics used to compare the two forecast types. The final section will explain the analysis technique used to evaluate the results.

The Data Base

The data used in this thesis consists of the following information: year, day, Ap (observed), F10.7 (observed), Ap forecasts (made 1, 2 and 3 days previously) and F10.7 forecasts (made 1, 2 and 3 days previously). The period of analysis is from 1 Jan 1971 through 29 Apr 1984, a total of 4868 days. This period was selected to provide values covering a period slightly larger than the 11 year solar cycle.

The bulk of the data base was obtained from the Environmental Simulations Branch, Detachment 3, HQ AWS, Sunnyvale AFS, California. It will be refered to as the Sunnyvale data. This data was received on a magnetic tape and contained: Ap (obs) and F10.7 (obs) values from 1 Jan 1965 through 29 Apr 1984: Ap (fcst) values from 1 Jan 1975 through 29 Apr 1984; and F10.7 (fcst) values from 25 Apr 1975 through 30 Apr 1984.

The missing forecast data was obtained from Operating Location B (OL-B), AFGWC, Boulder, Colorado. It will be referred to as the Boulder data. This data was received as hard copy tabulations. Each page contained one month's observed and forecast values for one index. Ap data was

ever, an observation system to directly monitor EUV and energetic particle input does not exist and is not likely to exist in the near future. Therefore, real-time users must rely on these admittedly inadequate indices. Point forecasts are made of these indices and are released jointly by AFGWC and the SESC. Although the SESC aknowledges the need for these forecasts to be verified, there is an apparent lack of an active verification program, particularly with respect to statistical analysis of forecaster skill.

as feedback to the user and the forecaster.

3. A summary of all currently used verification techniques and their advantages is needed.

4. It is imperative that all published forecast accuracy statements include: (a) a clear statement of exactly what is being forecast, (b) the climatology during the forecast period, and (c) a clear statement of exactly how the verification was performed (Smith, 1979:431).

Compared to the number of reports and papers written by meteorologists on the topic, it appears that there is a lack of standardization and direction and that there is definite room for improvement.

The 1982 Workshop on Satellite Drag showed little has been done to improve the verification situation in the US. The paper by Joselyn (1982a) titled "SESC Geomagnetic Predictions" included a brief discussion on verification. It described the "percentage of hits" score and explained that records were kept explaining reasons for over- and under-forecasts. Overforecasts result from major flares which produce no strong geomagnetic activity. Under forecasts were blamed on flares, filiment disappearances, and combinations of coronal holes, small flares, and/or disappearing solar filaments (Joselyn, 1982a:142).

Summary

From this review, the following points are worth emphasizing. Ap and F10.7 are indices used in atmospheric density models because they are indicators of upper atmospheric heating. They do not actually measure the elements involved in the heating, which is their primary disadvantage. How-

evaluated using a modified Brier P-Score (Heckman, 1979:344). This score is mathematically proper using the method of Murphy and Epstein and does measure the skill of the forecaster.

The monthly verification summaries for Ap and F10.7 also contain the percentage of "hits" for each variable. A "hit" is defined as a forecast which falls within +/- 10 units of the verifying observations. The percentage is computed as the number of hits divided by the number of forecasts. The limit of 10 units is a customer defined number which indicates the sensitivity of the atmospheric density models to those values (Eis, 1984; Roehrick, 1984). In other words, a forecast which is not exact but within this limit will not adversely affect the output from the density model (Procheska, 1984;4). A significant error is the term used to define a forecast which is not a hit. The percentage of significant errors can also be calculated. It is equal to one minus the percentage of hits.

During the Solar-Terrestrial Predictions Workshop, a forecaster's meeting was held "to discuss the early progress of each of the working groups" (Smith, 1979:428). Forecast verification was one topic at the meeting and the following comments reflect the status of the space environment forecast evaluation programs:

> 1. There is a need for more utilization of statistical analysis to develop and verify forecasting techniques.
>
> 2. Verification may indicate which variables are the good predictors as well as indicate the quality of the forecast. Verification is valuable

It is important to understand the difference between skill and accuracy. The distinction is important for the choice among the many scores available to the evaluator: some scores measure accuracy, while some are designed to measure skill. Accuracy is a measure of the size of the error, the difference between what is forecast and what is subsequently observed. Skill is a measure of the forecaster's knowledge and ability to predict future events or conditions. An accurate forecast may not be skillful and vice versa. For example, in certain stable climates, like a southern California summer, persistence may be an accurate forecast, but there is no skill involved. On the other hand, a forecaster who is able to predict the rare occurrence of rain during the summer in southern California, over and above the climatic probability of such an event given a set of initial conditions, is demonstrating his or her skill.

In the space environment field, forecasts are generally verified using measures of accuracy rather than skill. AWS regulation 178-1 offers guidelines to verify solar and geophysical forecasts. These guidelines instruct AFGWC to compute mean daily error, root mean square error and the ratios of MDE and RMSE to the standard deviations of the observed values for F10.7 cm solar radio flux and Ap geomagnetic forecasts (Dept of the Air Force, 1983:3-2 to 3-3). These are measures of accuracy, not skill (Abraham and Ledolter, 1983:374).

An exception is the flare forecasts jointly issued by SESC and AFGWC. These are probability forecasts and are

Silvert discusses the need for forecasting services to structure their forecasts based on the needs of the customer and then to derive an appropriate scoring system/evaluation technique and not vice versa. However, he says that if a scoring rule cannot be devised than the forecast is useless. Given a scoring system, he goes on to say that the success function must be normalized by accounting for climatology. More weight must go to the correct forecast of an unlikely event. This normalization gives flexibility to scoring types of forecasts which are useful to clients but hard to evaluate. The normalization procedure he proposes measures the "effectiveness" of a forecast. An ineffective prediction implies that climatology would be a better predictor (Silvert, 1980:146). The method appears to be most useful for probability forecasts although it is apparently applicable for point and categorical forecasts also (Silvert, 1980:149).

The paper by Gringorten, et al, presents a scoring method which measures the skill rather than the accuracy of a forecast. Skill is defined as the forecaster's ability to "recognize and measure the probability of departure of the future event" based on the cumulative climatic frequency of the forecast value (Gringorten, et al, 1980:189). The score is non-mathmatically proper and is for measuring exact forecasts, not categorical or probabilistic forecasts. This method can be statistically compared with a random, unskilled prediction using a Chi-Square or Binomial test of significance (Gringorten, et al, 1980:191).

average temperature (below normal, normal, above normal). Climatology is used to divide the temperature scale into three equally likely ranges. The random forecast would therefore have a 33% chance of being correct while the climatology prediction would always be for normal temperatures. A persistence forecast would just use this month's mean temperature. If the forecast was used by an energy company to determine next month's heating needs and the purpose was to determine whether it was worth the forecaster's time to prepare the prediction, it might be necessary to compare all four forecasts (ie. the forecaster's and the three control predictions). Ideally the forecast should beat climatology, but if persistence also beats climatology, or if a random guess does just as well, the forecast service is not justified. If the purpose of the prediction was to identify meteorological variables important in long-term forecasting, then a random forecast would be less useful than persistence since persistence forecasts could be analyzed given current meteorological conditions.

The area of choosing a control forecast is an active one in the weather business. The World Meteorological Organization (WMO) held a symposium in 1980 on Probabilistic and Statistical Methods in Weather Forecasting. During the sessions on Model Verification and Forecast Evaluation, no less than six of the fifteen papers presented were about applying one of the three types of control forecast, with the emphasis on climatic probabilities. Of particular interest are the papers by Silvert, and Gringorten, et al.

measures which are meaningful depending on the purpose and attributes of the verification.

Since many purposes of evaluation require a comparison between forecasts, it is necessary to have two forecasts. This is not a problem when comparing skill between different forecasters or when comparing different forecast methods, but it does become a problem when the data consists of just one set of forecasts and the evaluator is interested in how good this set of forecasts is. A control forecast is needed for comparison.

There are three types of control forecasts which can be used as a standard for comparison; random, persistence and climatology (Brier and Allen, 1951:846). Random, or chance forecasts can be generated from a uniform distribution for point and probability forecasts and from a contingency table for categorical forecasts. Persistence forecasts simply use the current condition as the next forecast. They are best for categorical and point forecasts. Climatology forecasts require a previous climatological data base or the creation of climatic probabilities from the observed data base (ie. a frequency distribution). This type of control forecast can be used for all three types of forecast. Control forecasts are also called unskilled forecasts since they may be produced without any scientific or technical knowledge.

Naturally, the choice of control forecast depends on the purpose of the verification and the use made of the forecast. For example, all three control forecasts could be used to make a categorical prediction of the next month's

the qualitative sense (Gringorten, et al, 1980:189-193). It will be discussed later.

In an another article, Murphy and Epstein (1967a) discuss the forecast evaluation process as a series of four steps where "elements" of the process are identified. Evaluation is defined to be synonomous with verification. It is worth describing the steps to note the similarity with the above criteria. The first step identifies the purposes of the evaluation which in turn defines the "form" of evaluation to use in the following steps. Step two identifies and defines attributes of the predictions. Attributes are desirable properties of the forecast for the purpose(s) chosen in step one. Validity and bias are two attributes discussed which are based on the association between predictions and observations on an individual and collective basis respectively. Quantitative measures are formulated in step three to determine which predictions possess a specific attribute. The measures discussed all relate to probabilistic predictions. Mean error and mean square error are mentioned as measures of validity and bias. Finally, statistical tests are developed to draw inferences from the results of the measures applied to predictions.

Murphy and Epstein argue that failure to apply this process has contributed to the controversy surrounding the subject of verification (Murphy and Epstein, 1967a:755; Heckman, 1979:344; Brier and Allen, 1951:841). This controversy is related to attempts to define a "best" measure of forecast accuracy. There is no single "best" measure, only

will be between 83 and 87), or a probability (there's an 80% chance today's high will be above 85). Objectivity eliminates the element of judgment to influence the comparison between the forecast and subsequent observation. It is essential to meet this criterion, which answers the question "How good is good?" An economist who predicts a "strong" recovery, a weather forecaster who predicts a "cold" night, or a market analysis who predicts a "good" year for sales would all have a hard time verifying their forecast unless they quantified strong, cold, and good.

The second criterion is to specify the purpose of the verification. Brier and Allen (1951:843) suggest stating the purpose of verification as a hypothesis. This allows easier selection of a scoring system to satisfy the purpose. It will also leave no doubt as to what action is indicated by the numerical value of the verification score.

Finally, the selected verification scheme should not influence the forecaster's predictions. Ideally, a forecast should represent the forecaster's true belief about what will happen. Knowledge of the verification score may influence the forecaster's decision. Murphy and Epstein use the term "proper" to define a scoring method which prevents the forecaster from "hedging" a forecast in order to improve his or her score (Murphy and Epstein, 1967b:1002-1004). They have devised a mathematical definition to determine whether a score is proper. Unfortunately, this definition only applies to probability forecasts. There is at least one score which applies to point forecasts which is "proper" in

Verification is a check of the accuracy of a forecast. The act of verifying a forecast is an important one. Since the objective of forecasting is to minimize forecast error, verification is a necessary final step in determining how good the forecast was. This basic objective of verification underlies the variety of purposes of forecast verification.

There are many purposes of forecast verification. When national weather services first began operating, forecasts were verified to justify the service's existance (Brier and Allen, 1951:841). A business executive may use economic predictions to determine marketing strategies. Verification might therefore be able to place an economic or utility value on the forecast. A weather organization uses verification to compare the relative skill of different forecasters. Many types of forecasts are verified to determine whether there has been an increase in accuracy over a period of time or between different time periods. Another purpose is to determine possible sources of forecast error and to perhaps identify variables which are good and bad predictors. Finally, verification is necessary to compare different forecast methods (Brier and Allen, 1951:481-482; Smith, 1979:431; Heckman, 1979:344; Murphy and Epstein, 1967b:748-749).

Brier and Allen (1951:842-843), in an early paper on forecast verification, discuss three criteria which should be met by a verification technique. The first is objectivity. This means the forecast should be stated as a point (the high today will be 85 degrees), a category (the high

EUV flux and F10.7, but it is not perfect. During the Solar-Terrestrial Predictions Workshop, a working group met to discuss user requirements of predictions for spacecraft applications. A neutral atmosphere subgroup further defined user requirements, current status of models and predictive techniques and recommendations for research and improvement. They concluded there was a critical need for more accurate predictions of F10.7 and Ap models. However, since F10.7 does not truly characterize the physical heating of the atmosphere, they recommended direct monitoring of EUV from space to aid in the future development of more accuarte models (Vampola, et al, 1979:13-16). A similar recommendation was made by an Air Force Scientific Advisory Board which examined existing density models with respect to predicting satellite ephemerides (Prochaska, 1984:5-6).

Thus, like the Ap magnetic index, it appears the 10.7 centimeter solar flux is not the best index. Because of the historical data base and economic availability of these indices, existing models would best be served by the most accurate F10.7 and Ap forecasts; more realistic indicators do not exist and would be expensive to obtain.

Verification

> ...the entire process of comparing the predicted weather with the actual weather, utilizing the data so obtained to produce one or more indices or scores and then interpreting these scores by comparing them with some standard depending upon the purpose to be served by the verification.
> (Brier and Allen, 1951:841)

regression equations were first developed for the Air Force in 1966. According to Prochaska (1984:4), the original equations were subsequently revised although information is not available as to when and why. The original and revised equations are listed in the Appendix. These qualitative predictions are then subjectively modified by forecasters to account for the effects of the current active region situation (Prochaska, 1984:4-5).

Separate forecasts are made by AFGWC and SESC personnel and then compared. Differences are settled before a single joint forecast for the next 1,2 and 3 days is issued (Ashton, 1984). Heckman (1979:341) briefly discusses the SESC subjective modification based on additional solar observations. Prochaska (1984:5) claims that modifications are made based on "...numerous, poorly-defined measures of solar variations" and raises the question that time spent forecasting may be better spent if the modified predictions are not better than the regression output.

Prochaska analyzed four years of data (July 1979-June 1983), comparing both regression forecasts with the forecaster's predictions. He concluded that the revised regression equations produced the most significant errors and that the slightly fewer number of significant errors produced by the forecasters over the original equations did not justify the forecaster's time spent in making the predictions. He recommended that the original equations be used as predictions without forecaster modification (Prochaska, 1984:16-19).

As mentioned earlier, there is a correlation between

operations person may occasionally fail to recognize the altered Ap value, hence an error is made in recording the Ap values. These errors are generally quite small, certainly no greater than 10 units from the actual value.

The second type of error may result by the key punch operator. The operator may misread a value or enter an incorrect number. The errors can be potentially large in this case. For example, if the F10.7 observed value is 222, the operator could hit an adjacent key without noticing it, and enter 111, or else could hit the space bar too many times and enter 22. These are admittedly extreme examples but ones that are entirely possible. Errors of this sort are made and can slip by unnoticed.

The above discussion is included because errors were found in the data base. These errors were discovered during initial examinations of the frequency distributions of the observed and forecast values and the forecast errors. While corrections were not documented, the number of corrections made numbered less than 10, primarily in the F10.7 data set. For example, the original frequency distribution of the second day F10.7 forecast contained one value of 22, clearly an unrealistic number. A systematic search of the data revealed that on 11 Jan 1977 a forecast of 22 was recorded for the second day while forecasts of 71 and 72 were make the first and third days while the observed value was 74. The erroneous observation was attributed to key punch operator error and corrected to a value of 72.

As previously mentioned, hard copy data was received

from OL-B in Boulder. Approximately 1400 lines of forecasts were entered by the author. Errors of the type attributed to the keypunch operator were made. Significant errors were corrected. However, small errors on a scale similar to Ap errors described above could have eluded the author's quality control review of this additional data.

It is nearly impossible to find instances where the "ones" or "tens" column is off by a single digit. However, the data base consists of 4865 lines, so it can be safely assumed that small errors of this sort will not adversely affect the results of this research. For example, the mean of F10.7 (2 day forecasts) increased from 129.881 to 129.892, a difference of .011 or .0085% of the mean. It would require a significant amount of errors as large as this to affect the analysis.

After the data were entered and all obvious errors removed, a program was run which created the persistence "forecasts." This involved taking the column of observed values, copying it to another column and placing the first element in the second, third or fourth row to create a 1, 2 or 3 day persistence forecast. Since observations were not available for the last three days of 1970 to create the persistence forecasts for 1 Jan 1971, the original data set was reduced by three days to yield 4865 days of forecasts to be verified.

The SESC forecasts had already been entered in a similar manner so that the verifying observation is on the same line as the three forecasts. It then becomes an elementary

operation to take the difference between forecast and observation to determine the forecast error and other statistics.

There is one more point to make with regard to the data base which concerns two assumptions made by the author. These assumptions deal with data collection methods and forecaster personnel. Although the F10.7 value has been observed from the same location for over thirty years, the same cannot be said about the magnetometer observatories used by the Air Force in calculating their real-time Ap values. A change was made in the five station magnetometer network in the late 1970's resulting in the establishment of an observatory at Upper Heyford AFB, England (Dye, 1984; Patterson, 1984). Although one might expect this new European data source to affect the Ap values calculated by AFGWC, a report by Dandekar (1982:8-9) revealed no significant differences between the periods when changes occurred. The assumption is therefore made that the data consists of a single continuous set. In other words, it is assumed that network operations changes have no effect on the production of AFGWC Ap values.

A second assumption is made regarding forecaster skill. The predictions have been made by a number of different forecasters with various levels of skill and scientific knowledge. If a forecaster had the ability to make predictions which were more accurate than forecasts of others, this factor would somehow have to be accounted for during the analysis since the total results would indicate a quality of the forecasts which is no longer present.

A question of this sort was posed to AWS personnel formerly associated with space environmental forecasting. Patterson (1984) is of the opinion that the experience levels were superior in the late 1970's than today and in general forecaster skill was better in the last decade compared to the present. His reasoning is that the current selection process arbitrarily chooses forecasters, whereas before the forecast team was "hand picked." Townsend (1984) makes the point that F10.7 prediction methods have been unchanged for 14 years with no apparent change in skill. There has been an improvement in the theory of how Ap behaves during this period although prediction techniques remain the same, ie., highly subjective. It is his feeling that the forecast accuracy would not reflect this additional knowledge. The author has decided to assume that any differences in forecaster skill levels would not significantly affect the results of this analysis.

Statistics for Comparison

Numerous scores and statistics are available for evaluating forecasts. Accuracy, bias, and skill are three attributes a score may be able to measure. Accuracy will be defined as a measure of the absolute amount of forecast error. Bias is a form of accuracy but is differentiated from it such that bias can measure whether the forecast(s) tend to be higher or lower than the verifying observation (Abraham and Ledolter, 1983:372-374). Skill is a measure of ability of a forecaster to make predictions which are better

than those which could be obtained using an unthinking method. This section will describe the various statistics and methods used to compare the data.

Forecast Error. The most basic statistic necessary for verifying any forecast is the forecast error. This is defined as the forecast value minus the observed value. A positive value is interpreted as an overforecast while a negative error is an underforecast. The absolute forecast error or absolute error is simply the absolute value of the forecast error. The absolute error will be used later to determine which forecast method was closest to the verifying observation. Taking the mean of the forecast errors can show forecaster bias while the mean of the absolute values of the errors is a measure of accuracy.

Frequency Table. A simple way for the evaluator to get a "feel" for a set of forecasts is to compile a frequency table of the errors. A frequency table is a tabulation of all the discrete error sizes with additional columns for the number of times a particular error was made, the absolute frequency that error was obtained and the cumulative frequency of the errors. Absolute frequency is defined as:

$$\text{Abs Freq} = N_{ei}/N \qquad (1)$$

where

N_{ei} = number of forecast errors of size i

N = total number of forecasts

Cumulative frequency is the sum of the absolute frequencies

and will equal unity at the last row of error size. A table of this sort is useful because it shows the range of errors and the number and percentage of errors that are particularly good or bad. If all the errors are divided by the standard deviation of the observed values, a process Abraham and Ledolter (1983:373) call "standardizing the errors," it is possible to visually check whether the errors are normally distributed.

Root Mean Square Error. One popular verification statistic appropriate for numerical forecasts is the root mean square error (RMSE). The RMSE is given by the formula:

$$RMSE = [\sum(Fi - Oi)^{2} / N]^{.5} \qquad (3)$$

where

Fi = i'th forecast value

Oi = i'th observed value

i = 1,2,3,...,N

The sum is taken of i from 1 to N

The RMSE is a measure of the accuracy of a forecast, it cannot show bias. This statistic gives greater weight to larger errors due to the squaring of the errors, therefore the better the forecasts, the smaller the RMSE. While the mean error statistic shows the average of all the errors, RMSE indicates the typical amount of error of a forecast. In this respect, it is similar to the standard deviation, which may be considered "the distance of a typical measurement from the mean" (Prochaska, 1984:7).

Brier and Allen (1951:844-845) warn that this scoring method allows a forecaster to hedge by choosing the middle value of a range when he or she is uncertain of what end-point the value will fall on. For example, assume Ap is being forecast and a major storm is expected but the forecaster is uncertain when it will arrive. If the storm arrives, Ap will jump from its current value of 15 to 50. If the storm doesn't arrive Ap will only increase to 20. The cautious forecaster, concerned with maximizing the score over making an honest forecast, would do well to choose 35 as a forecast. This value reduces the forecaster's maximum error even though he or she is sure that a value of 35 will be incorrect. A forecaster may hedge in a similar manner if mean error or absolute mean error is the scoring method (Brier and Allen, 1951:843).

In spite of this deficiency, AWS uses a modified RMSE and mean error as the basis for their monthly verification of F10.7 and Ap (Dept of the Air Force, 1973:3-2 to 3-3). These statistics are modified by dividing by the standard deviation of the observed values. Although supporting documentation has not been found, it is believed that the ratio is taken to account for the variance of the month's observations (Ashton, 1984; Schleher, 1984).

In this respect, the ratio may be regarded as a standardized value. The advantage to this is the new scores are now more comparable between periods. The reasoning is that during periods of highly variable solar activity the RMSE and mean error would naturally be larger than during periods

of quiet activity. Comparisons between the two periods would be biased towards the smaller values associated with the quiet periods. By taking the ratio of RMSE to standard deviation, the observed variability is removed and a more realistic comparison can be made.

There is still some controversy with this ratio statistic. Ashton (1984) raises the following argument: the periods of high variability are generally during the declining phase of the solar cycle when much of the activity is associated with recurrent activity on the sun. This variability then becomes predictable since the amplitude of the recurrent cycle is fairly stable. In this respect, the ratio would tend to favor verification periods when recurrence is the most likely cause for variations in the observations. These periods may then tend to have lower ratio values than periods when the variablilty is small but observations are punctuated by large, unpredictable storms such as those during solar maximum.

It should be noted that this whole ratio issue is relevant only when comparing different periods of forecasts. When evaluating two forecast methods conducted over the same period, it is irrelevant whether RMSE or RMSE/SD is used.

Significant Errors. A significant error is the term used by Prochaska (1984:4) to define an Ap of F10.7 forecast error which is greater than 10 units. The threshold of 10 has been established by NORAD and Sunnyvale as a sensitivity limit of the forecasts (Eis, 1984; Roehrick, 1984). Errors less than or equal to 10 units will not significantly affect

the density and drag models.

A scoring method used by AWS in their verification procedures for Ap and F10.7 is the monthly production of the percentage of hits. This score is not defined in AWSR 178-1 but falls under the guideline to "establish standards based on the state of the art and customer requirements" (Dept of the Air Force, 1983:2-3). A hit is a forecast whose error is not significant. The percentage of hits can be used to compare forecast techniques within a period, but, like RMSE, comparing between periods can be misleading if there is a tendency to make more significant errors during active periods.

This score, while informative, has a long and controversial history. Two papers from the early fifties (Brier and Allen, 1951:846; Gringorten, 1951:280), point out that this score is "meaningless" when it is not compared with some type of "blind" forecast. The percentage of hits has been incorrectly described to measure a forecaster's skill. This score can only show skill when compared with random, persistence, or climatological forecasts; examining the number of hits in excess of those obtainable by one of the above blind forecasts. A score attributed to Heidke has been developed which incorporates such a comparison (Gringorten, 1951:280).

Additionally, this score is subject to hedging in a similar, if not worse manner than RMSE. Forecasters can protect their score by not making any extreme forecasts (Schleher, 1984). Continuing the above Ap storm forecast example, a hedge against this score would be to forecast 30.

This forecast will not produce a significant error if the storm does not occur although the forecaster may reason it does have a certain amount of warning associated with it.

The scores discussed in this section are useful for describing certain properties of the forecasts like bias and accuracy. They are easy to understand and easy to use when comparing forecast methods within periods and forecasts between periods. For these reasons, these scores will be called descriptive scores or statistics.

Analysis Technique

One of the faults with the descriptive scores is that they cannot be used to make statistically effective comparisons between two forecast types. The forecaster RMSE may be lower than persistence RMSE but there is no way to determine if this difference is statistically significant. The forecast verification field has many descriptive scores but few statistically usable ones, hereafter called comparative scores. This section will describe the test used to statistically compare the two forecasts.

The sign test is the technique of chioce for determining whether the forecasters exhibit skill when compared to persistence. This test was suggested by Boehm (1984) who said that the test is appropriate for the task at hand and that it had sufficient power in its ability to adequately distinguish between the two forecast types.

The sign test is a nonparametric test which is used with paired data to test if one random variable in the pair

tends to be larger than the other random variable. Therefore, the data must be at least measured on an ordinal scale, that is a scale where one can order the data elements based on their relative size. This feature makes the sign test particularly attractive for checking the number of significant errors made between the forecasts. The sign test will also be applied to the absolute differences of the two forecast errors, although it does not distinguish between the size of the errors.

This is a disadvantage to using the sign test since information of the data, the difference in size of the errors, is not used. Alternative nonparametric tests (Wilcoxon signed rank test) and parametric tests (paired t-test) are available which have more power and make use of the size of the error differences. However, they require additional assumptions to be made about the distribution of the differences, namely a symmetric or normal error distribution. If these assumptions are made when they are not in fact true, the test results will be biased and show a tendency to reject the null hypothesis when it is true. By using the less powerful test, the analyst will not run the risk of making such an error and can confidently reject the null hypothesis. The data was analyzed with respect to these assumptions and the author felt that the difference between the mean and the median was large enough to affect the symmetry for the signed rank test assumption and that the frequency distribution of the errors was skewed enough to affect the normal assumption needed to use the paired t-

test. The following description of the test procedure is taken from Conover (1980;122-128). The significant error application is discussed first.

The first step requires assigning one of two classification values to each datum pair (Fi, Pi) where Fi is the forecaster error and Pi is the persistence error. If either error is significant, that particular element is assigned a value of "1." If the error is less than or equal to 10 (ie. a "good" forecast which will not adversely affect the density model), the element is assigned a value of "0." There are four combinations which each pair may fall into: either both are good (0,0), both are bad (1,1), or one is good while the other is bad (0,1), (1,0). This test is interested in the total number of (0,1) and (1,0) pairs. The number of times the forecaster was good while persistence was bad (0,1) will be defined as Nf. The number of times persistence was good while the forecaster made a significant error (1,0) will be defined as Np.

If there was no difference between the forecaster and persistence, then one would expect Nf to equal Np. Preferably, the forecaster exhibits some skill compared to persistence and Nf would be greater than Np. The null and alternate hypotheses may now be stated:

$$H: \; Nf \leq Np \quad \text{versus} \quad A: \; Nf > Np$$

The value of Nf can be considered the test statistic which determines which hypothesis to accept. Let N be the sum of Nf and Np (note that ties are disregarded, it doesn't

matter when both are good or bad). The decision rule is to reject H when Nf is greater than (N - t) where t is found in a binomial table with p = 1/2 and N at the appropriate significance level, alpha. When N is greater than 30, the normal approximation may be used to find t from the equation

$$t = .5[N + z(N)^{.5}] \tag{3}$$

where z is obtained from a table of normal probability values at the desired alpha level.

The significance level chosen for this thesis is .05. This level was selected from the working group report by Vampola, et al (1979:4), which recommended that users of Ap and F10.7 predictions wanted a 5% level of accuracy. The Statistical Package for the Social Sciences (SPSS) (Nie, et al 1975) is used for the calculations. Therefore the p-values will be listed with the understanding that the null hypothesis will be rejected when the p-value is less than 0.05. The proper interpretation of the p-value is that there is a "one minus the p-value" probability that Nf is larger than Np. Alternately, there is only a "p-value" percent chance of rejecting H when H is in fact true. In either case, rejection of H leads to the conclusion that the forecasters performed better than persistence during that particular period.

The application of the sign test to the absolute error data is identical except that Nf is the number of times the absolute error of the forecaster was less than the absolute error of persistence, Np is the total of the opposite situation occurrences and there is only one form of tie.

IV. Results and Analysis

This chapter is divided into three sections. Section one introduces the order of analysis, describes the various data tables and discusses the observed data values with respect to the period of the solar cycle which the data covers. Section two analyzes Ap results; section three analyzes F10.7 results.

Data Tables

Within each section, RMSE is analyzed first, significant errors are compared next and finally the results of the sign test for significant errors and absolute errors are presented. This order of discussion is followed for the one, two and three day forecasts separately. The scores for each forecast day are presented in a table which contains annual scores in addition to the total data base scores. The forecaster scores are abbreviated as fcst or fcs while persistence is abbreviated as pers or per.

Analysis of each table will include a comparison of the total scores for forecaster and persistence plus a comparison of unusual or significant annual scores, particularly with respect to the solar cycle. Each section will conclude with a comparison between the scores for different forecast days.

The RMSE tables include the standardized ratio of RMSE to the standard deviation of the observed values. This ratio is abbreviated as RMSE/SD. Keep in mind that this standardized score is used for comparison between periods

Table 4-8

Ap Significant Error Breakdown and Sign Test for

First Day of Predictions

Year	Good	Bad	Nf	Np	P-value
1971	262	32	38	30	.198
1972	278	32	34	22	.071
1973	236	47	44	38	.291
1974	221	51	47	46	.500
1975	231	54	43	37	.288
1976	283	28	32	23	.141
1977	277	36	34	18	.019
1978	233	50	50	32	.030
1979	258	43	23	41	.017
1980	289	23	23	31	.171
1981	228	46	48	43	.338
1982	203	45	58	59	.500
1983	221	47	52	45	.271
1984	73	14	15	18	.364
TOTAL	3293	548	541	483	.038

Table 4-7

Ap Significant Errors (in percent) for

Third Day of Predictions

Year	Total		Underforecast		Overforecast	
	Fcst	Pers	Fcst	Pers	Fcst	Pers
1971	20	31	14	15	6	16
1972	21	28	16	14	5	14
1973	30	43	23	21	7	22
1974	29	47	19	21	10	26
1975	29	46	17	22	12	24
1976	15	29	11	16	4	13
1977	21	25	12	13	9	12
1978	33	40	16	20	17	20
1979	25	28	12	15	13	13
1980	13	21	8	11	5	10
1981	26	31	15	15	11	16
1982	33	45	23	22	10	23
1983	28	45	18	22	10	23
1984	30	41	18	23	12	18
TOTAL	25	36	16	18	9	18

Table 4-6

Ap Significant Errors (in percent) for

Second Day of Predictions

Year	Total		Underforecast		Overforecast	
	Fcst	Pers	Fcst	Pers	Fcst	Pers
1971	18	17	13	13	5	14
1972	19	27	14	14	5	13
1973	27	38	21	18	6	20
1974	33	37	20	18	13	19
1975	29	40	16	20	13	20
1976	14	25	10	13	4	12
1977	19	22	12	11	7	11
1978	33	37	16	18	17	19
1979	26	27	12	14	14	13
1980	14	20	5	9	9	11
1981	27	32	13	16	14	16
1982	33	43	21	21	12	22
1983	27	43	18	22	9	21
1984	30	38	17	20	13	18
TOTAL	25	32	15	16	10	16

predicting sudden decreases in Ap than predicting the increases. In eight years, persistence had fewer underforecast errors, while the forecasters had fewer overforecast errors in 11 years. In six of those 19 years, the difference was at least four percent.

Moving to the second and third day predictions (Tables 4-6 and 4-7), a definite difference between forecast types becomes apparent. Forecasters make 7 and 11 percent fewer total errors and 6 and 10 percent fewer overforecast errors. Persistence still compares favorably on the amount of underforecast errors committed. One thing to note is these differences result primarily from an increase in the number of persistence errors. Forecaster error scores are almost constant between the second and third days and increase only by four percent from the first day scores. The worst years (i. e. highest percentage of errors) continue to be the ones when the geomagnetic field is most active.

Significant Error Sign Test Results. The significant error breakdown and sign test tables (Tables 4-8 to 4-10) provide considerable comparative information in addition to the sign test p-values. Recall from the last chapter that the integers listed in these tabluations are the number of days the paired data occurred in one of the four possible combinations: both good (under the heading "good"), both bad ("bad"), forecast good and persistence bad ("Nf"), and forecast bad and persistence good ("Np"). Also keep in mind that larger values are preferred for the good and Nf columns

Table 4-5

Ap Significant Errors (in percent) for

First Day of Predictions

Year	Total		Underforecast		Overforecast	
	Fcst	Pers	Fcst	Pers	Fcst	Pers
1971	17	19	12	10	5	9
1972	15	18	12	10	3	8
1973	23	25	18	13	5	12
1974	27	27	15	14	12	13
1975	25	27	14	15	11	12
1976	14	16	10	8	4	8
1977	15	19	9	9	6	10
1978	22	28	11	14	11	14
1979	23	18	9	9	14	9
1980	15	13	7	7	8	6
1981	24	26	12	14	12	12
1982	29	28	17	13	12	15
1983	25	27	16	13	9	14
1984	27	25	15	12	12	15
TOTAL	21	22	12	11	9	11

in seven instances. While the differences are generally quite small, there is an indication that the three day forecast was a little more accurate than the two day forecast. One explanation for this is if a solar flare is going to affect geomagnetic activity, there is usually a three day traveling period for the particles to be carried out in the solar wind before hitting earth.

Significant Error Comparison. Table 4-5 lists the percentage of significant errors for the first prediction day. The first thing to notice is that there is only a one percent difference between the two total scores. This implies no preference for either the forecaster or persistence prediction. However, in nine of the fourteen years, the forecaster percentage was less while the persistence percentage was less in only four years. This implies a trend toward a lower number of significant errors for the forecaster. Note in 1978, a year of moderately high geomagnetic activity, forecasters had six percent fewer significant errors, while in 1979, persistence beat the forecasters by 5 percent. Overall, the total percentage seems surprisingly high, i.e., one out of five one day forecasts will be significantly wrong. Fewer errors are made during periods of relatively low activity, while many more errors are made when geomagnetic activity is high. This is based on comparing the lowest percent of errors with the low mean Ap values and the highest error scores with the high Ap mean values.

When one looks at the numbers for under- and over-forecasting, it appears the forecasters do a much better job

Table 4-4

Ap RMSE and RMSE/SD Values for

Third Day Predictions

Year	RMSE(fcs)	RMSE(per)	RMSE/SD(fcs)	RMSE(per)
1971	11.9375	16.5072	1.0331	1.4286
1972	16.5996	22.9077	1.0032	1.3844
1973	15.0263	19.1179	.9928	1.2631
1974	14.7802	20.2881	.9808	1.3463
1975	11.7323	17.3611	.9298	1.3760
1976	10.1479	15.4276	.8474	1.2883
1977	9.8544	13.2229	.9915	1.3305
1978	15.2802	21.5025	.9659	1.3593
1979	11.2796	12.7284	1.1049	1.2468
1980	7.6435	9.1971	1.1242	1.3529
1981	14.2665	15.6253	1.1681	1.2794
1982	18.0053	23.3452	1.0222	1.3253
1983	13.9871	19.3705	1.0437	1.4454
1984	14.1980	18.2852	.9829	1.2658
TOTAL	13.4434	17.9027	.9775	1.3017

This page intentionally left blank.

RMSE(pers) for every year except 1980 which was the unusual year where the Ap observed mean was the lowest. Since a RMSE/SD ratio less than one means the RMSE was less than the standard deviation of the observed values, it is interesting that the forecaster RMSE scores were less than the standard deviation for all but three years (1971, 1979, 1980) while persistence RMSE scores were greater than the standard deviation for all but three years (1972, 1973, 1975). There appears to be a tendancy for the forecaster RMSE/SD values to be smaller during the early to mid-seventies which may support a hypothesis that forecaster skill was better before 1979; however, there is a similar trend in the persistence RMSE/SD values.

Table 4-3 gives the same scores for the second day of predictions. The difference between total RMSE scores has increased to 3.3152. Persistence has a lower RMSE in 1981 this time. The forecast ratio was less than unity eight years for the two day predictions, a decrease in three years over the one day predictions. No persistence ratios were less than one any more.

Table 4-4 reveals the third day of predictions show the RMSE difference increasing to 4.4593 while all forecaster scores are less than persistence scores. Only seven forecaster RMSE/SD scores are less than one. One interesting point is that the total forecaster ratio score for the third forecast day is less than that score for the second forecast day. An annual comparison of these ratio scores between forecast days reveals that the third day prediction was less

Table 4-2

Ap RMSE and RMSE/SD Values for

First Day of Predictions

Year	RMSE(fcs)	RMSE(per)	RMSE/SD(fcs)	RMSE/SD(per)
1971	12.2495	13.7256	1.0601	1.1878
1972	13.3598	15.8018	.8074	.9555
1973	13.5372	13.7811	.8944	.9105
1974	14.6353	15.3980	.9712	1.0218
1975	10.6465	12.3268	.8438	.9770
1976	10.4904	12.9442	.8760	1.0809
1977	9.1408	10.3555	.9197	1.0420
1978	12.8156	17.0298	.8101	1.0765
1979	11.1408	11.5827	1.0913	1.1346
1980	7.9078	7.3258	1.1630	1.0774
1981	11.6940	12.8893	.9575	1.0554
1982	16.4562	18.0628	.9342	1.0222
1983	12.5975	14.8233	.9400	1.1061
1984	12.7459	14.6233	.8824	1.0123
TOTAL	12.2594	13.8375	.8914	1.0662

The previous solar maximum occured in late 1968 (Prochaska, et al, 1981:67); therefore this data base begins and ends during the declining phase of the solar cycle. This means that F10.7 should have its maximum values in 1979-1980 since solar flux closely follows the solar/sunspot cycle. Ap, on the other hand, should be large in the mid-seventies and again during the eighties since its most active period is "during the declining phase of each sunspot cycle... in the years just before the solar minimum" (Fraser-Smith, 1972:4211). It is reassuring to note, then, that the minimum mean annual F10.7 value occurs in 1976 while its maximum value is in 1981. The minimum Ap value of 11.00 occurs in 1980 which is somewhat confusing, although the next lowest annual average is in 1977; its maximum values occur in 1974 and 1982.

Note the size of the standard deviations. They also vary with respect to the solar cycle. More importantly, when the SDs are compared to their means, the Ap SD is almost as large as its mean. This is because magnetic storms cause very rapid increases in Ap values, frequently 40 to 80 units above the average value. Variations of this size will have a strong effect on the standard deviation values.

Ap analysis

RMSE and RMSE/SD comparison. Table 4-2 gives the RMSE and RMSE/SD values for the first day of Ap predictions. The bottom line shows that the RMSE of the forecaster is 1.5781 units less than the RMSE of the persistence "forecasts" for the total analysis period. In fact, RMSE(fcst) is less than

Table 4-1

Mean and Standard Deviation of Observed Values

Year	Ap		F10.7	
	Mean	SD	Mean	SD
1971	11.4696	11.5550	118.2403	20.5422
1972	12.6148	16.5469	120.8689	21.2245
1973	17.0329	15.1356	93.3342	13.2807
1974	19.4575	15.0695	86.8493	12.8827
1975	15.3808	12.6174	76.2219	8.2247
1976	12.6557	11.9756	73.4454	4.6624
1977	11.3397	9.9385	87.0603	11.1760
1978	16.4822	15.8196	144.5315	26.8390
1979	14.6575	10.2086	192.9671	33.6274
1980	11.0000	6.7993	199.9727	34.4953
1981	16.0027	12.2132	202.5370	37.0184
1982	21.4000	17.6149	175.8301	39.0426
1983	20.0986	13.4011	119.8932	21.2968
1984	20.4000	14.4451	126.3333	24.5416
TOTAL	15.4781	13.7528	130.0491	51.9171

gives the annual (or total) number of days when both forecasts were "good" (0,0). The next column is the number of ties when both were "bad" (1,1). These columns are provided for comparison even though their values were not used in the test. The third column is Nf, the number of days persistence was a significant error while the forecast was not (0,1). Next is the column of Np, the number of days persistence was good while the forecaster commited a significant error (1,0). The last column lists the p-values for the one-sided hypothesis test discussed in the last chapter. Recall that the null hypothesis should be rejected when the p-value is less than .05. Please note that occasionally Np is greater than Nf. When this occurs, the alternate hypothesis changes from Nf is greater than Np to Np is greater than Nf. The times when this happens will be noted in the text.

The absolute error sign test results are presented in a table of nine values, three sets of three. Each set lists Nf, Np and the p-value for each forecast day. The number of ties is unimportant in this case.

Table 4-1 is a listing of the observed means and standard deviations for each index, broken down into annual values and the total value. It gives an indication of the annual variability of the indices. It is appropriate to note here that solar minimum occured in June 1976 while the last solar maximum occured in September 1979. These dates are based on observed monthly sunspot numbers (Springer, 1982:107). Sunspot numbers are widely regarded as the primary criteria for determining the period of the solar cycle.

(year vs year or total vs year). The ratio does not give any additional information for comparisons between forecaster and persistence within a period.

The significant error tables each consist of six columns. The first two columns are the total percent of significant errors for forecaster and persistence. The next two columns give the percent of underforecast errors while the last two tables give the percent of overforecast errors. A significant underforecast error is defined as (fcst (or pers) - obs) < -10 and indicates that activity increased more than expected. A significant overforecast error is defined as (fcst (or pers) - obs) > 10 and indicates that activity declined more than expected. Obs is the abbreviation for the observed value which the forecast is being verified against.

The percentage of persistence significant errors should be about equally divided between under- and overforecasts, if one assumes an equal number of sharp increases and decreases. Comparison of under and overforecast percentages between forecaster and persistence can give information about the skill of forecasters predicting the beginning or end of storms or active events. For example, if the forecaster percentage for underforecasts is notably lower than the corresponding persistence percentage, one can infer the forecaster did a better job at predicting a sudden increase in the index.

The significant error sign test results are presented in tables of five columns. The first column after the year

Table 4-9

Ap Significant Error Breakdown and Sign Test for

Second Day of Predictions

Year	Good	Bad	Nf	Np	P-value
1971	243	43	56	20	.000
1972	241	41	55	29	.003
1973	194	69	71	31	.000
1974	184	73	61	47	.106
1975	181	66	79	39	.000
1976	256	33	60	17	.000
1977	252	36	45	32	.086
1978	178	70	65	52	.134
1979	220	47	53	45	.240
1980	260	21	53	32	.015
1981	199	51	67	48	.047
1982	157	68	88	52	.002
1983	184	73	84	24	.000
1984	59	20	26	15	.059
TOTAL	2808	711	863	483	.000

Table 4-10

Ap Significant Error Breakdown and Sign Test for

Third Day of Predictions

Year	Good	Bad	Nf	Np	P-value
1971	228	52	61	21	.000
1972	234	47	57	28	.001
1973	194	69	71	31	.000
1974	157	70	100	38	.000
1975	154	63	104	44	.000
1976	241	34	70	21	.000
1977	238	39	50	38	.121
1978	162	67	80	56	.025
1979	218	49	56	42	.095
1980	265	24	53	24	.001
1981	206	51	63	45	.051
1982	156	74	90	45	.000
1983	160	64	103	30	.000
1984	56	21	28	15	.034
TOTAL	2645	727	1001	492	.000

and smaller values for the bad and Np columns. The reasoning is that a forecaster does not want to make significant errors and therefore wants Nf to be larger than Np and to minimize the bad column. However, if the forecaster is going to have a bad prediction, it would be better to be bad when persistence is also bad which means increasing the bad values while decreasing the Np values. It is important to understand this to avoid confusion when relating these tables to the percent significant error tables, where lower forecast percentages are indicative of better forecaster performance.

The most significant aspect of these three tables comes from the bottom line of each. The p-value is less than .05, so the alternate hypothesis of a statistical advantage of forecasters over persistence can be accepted. The difference between Nf and Np is not that large for the first day prediction but increases dramatically by the third day. Note that the bad column is larger than either the Nf or Np columns on day one, while by day three, the relative order of the bad, Nf and Np columns has approached the preferred relationship described above.

When the sign test is applied to individual years, it is revealing to note that for the first day predictions, persistence performs equally well with the forecaster and even performs statistically better in 1979. The p-value here would accept the alternate hypothesis that Np is greater than Nf. Np was also larger than Nf in 1980. Nf is only statistically better in the years just after solar minimum, 1977 and 1978. During the years when Ap was at

high mean levels with large standard deviations (1974 and 1982) it is interesting to see that Nf and Np are almost equal, with many days when both forecasts committed significant errors.

On the second prediction day (Table 4-9), the skill of the forecaster becomes more apparent. There are five years when the null hypothesis cannot be rejected at the five percent significance level and five when the alternate can be accepted with 100 percent confidence. This trend continues into the third prediction day (Table 4-10) when only three years cannot claim to show forecaster skill (1981 is on the borderline). There are now seven years when the alternate may be accepted with 100 percent confidence.

Absolute Error Sign Test Results. The results of the sign test when applied to the absolute error scores for the paired data is presented in Table 4-11. This table has all three days of forecast results which makes for easier comparison between days. For this test application, it is desired to have larger values under the forecast column which imply more days when the absolute forecaster error was less than the absolute persistence error. Remember that if the persistence value is larger than the forecaster value then the listed p-value gives the probability of accepting the alternate hypothesis that persistence performed better than the forecaster.

That last reminder is important to keep in mind because it turns out that for the first forecast day, persistence outperformed the forecaster a total of 97 times over this

Table 4-11

Ap Absolute Error Sign Test

Year	Day 1			Day 2			Day 3		
	Nf	Np	P	Nf	Np	P	Nf	Np	P
1971	153	169	.202	190	144	.007	201	122	.000
1972	157	166	.328	181	143	.020	185	148	.025
1973	164	167	.456	200	128	.000	214	123	.000
1974	164	166	.478	189	148	.015	219	124	.000
1975	171	154	.188	218	117	.000	237	107	.000
1976	161	165	.434	200	132	.000	221	118	.000
1977	152	183	.051	172	166	.393	178	162	.208
1978	176	152	.102	194	149	.009	195	140	.002
1979	169	168	.500	167	171	.435	184	158	.088
1980	148	187	.029	169	169	.500	185	158	.097
1981	170	172	.479	180	161	.165	175	156	.161
1982	162	175	.257	206	142	.001	211	137	.000
1983	160	175	.222	216	124	.000	224	120	.000
1984	52	57	.351	61	51	.298	57	58	.500
TOTAL	2159	2256	.080	2543	1945	.000	2684	1831	.000

thirteen-plus year period. This was almost enough to statistically conclude with 95 percent confidence that persistence was indeed a better forecast. The forecaster was only closer to the observed values more often than persistence in 1975 and 1978.

By the second forecast day, the forecaster once again begins to show skill in making more accurate forecasts. The years 1979 and 1980 (solar maximum) are the only ones where the persistence forecasts equal or exceed the forecaster predictions. The forecasters did very well in 1971 through 1978 with the exception of 1977 when geomagnetic activity was at its lowest level during that seven year period.

Day three of the predictions show an increase in forecaster ability with the exception of 1981 and 1984. There is hardly any improvement in 1981 and actually a decrease in the number of better predictions by the forecasters in 1984. The total number of more accurate forecasts by GWC has increased to 853 however.

Before examining the F10.7 scores and test results it is worth summarizing the Ap analysis. In general, the forecasters did not drastically outperform persistence on the first forecast day. However their skill was apparent on the second and third forecast days. The forecasters did tend to make fewer significant underforecast errors which implies an ability to predict the end of a disturbance better than their ability to predict the start of a disturbance. During solar maximum and shortly afterwords, both types of forecasts performed worse than the years around

solar minimum, although 1980 was an unusual year for geomagnetic activity when the observed mean and standard deviation took an unusual dip. In 1980, persistence often beat the forecaster, especially on the first forecast day. This is reasonable since a low standard deviation value would tend to restrict the day to day fluctuations of Ap. In 1977, the other low activity year, persistence performed similarly. Finally, it appeared that as a group the first six years of the forecasts outperformed the remaining years. This was during the slow decline in activity after the 1969 solar maximum and probably reflects better predictability during this phase of the solar cycle, although an argument may be made for more skilled forecasters then versus now.

A few words should be said with respect to the over/under forecast errors and the physics of the solar activity/geomagnetic activity relationship. Solar observations of flares and active regions are used in making Ap forecasts, particularly the regions where the activity is occurring. Activity on certain parts of the solar disk will not affect earth. The movement of an active region out of the area where it can produce geomagnetic disturbances aids in predicting a decline in Ap values. However, even when active regions and flares are in locations where their particles may hit the earth, geomagnetic activity is not guaranteed to occur. Hence the difficulty in predicting the onset of geomagnetic storms. Additionally, there is the problem of timing storm onset. Generally particles take three days in the solar wind, but not always.

F10.7 Analysis

RMSE and RMSE/SD Comparisons. Table 4-12 lists the F10.7 RMSE data. The difference between forecaster and persistence is .7841, a 12% improvement over persistence. Forecaster scores are lower in 10 of the years and are generally not much worse than persistence in the remaining four years (1971, 1972, 1975, 1976), with the exception of 1971 when the difference is 1.6758 in favor of persistence. The annual values appear to be correlated with the solar period exhibited by the mean annual values of F10.7 (see Table 4-1). The minimum scores for both forecaster and persistence occur in 1976, the year of solar minimum. The largest persistence RMSE score is in 1982, which is the year of the largest observed standard deviation. The forecaster's maximum score is in 1980, the year after sunspot maximum and the year before the observed F10.7 maximum mean value.

The columns of ratio data offer some insight into the quality of the F10.7 forecasts. Except for 1976 and 1984, the forecaster ratios are all very close to .25 or .26. This would imply that the forecast quality remains fairly constant over the solar cycle. One exception is 1976, the year of solar minimum. For this year, the observed standard deviation is very low, one half the amount of the next smallest standard deviation and almost one tenth as small as the year with the largest standard deviation. It appears the forecasters predicted for more variation than actually occurred, resulting in their largest ratio score, even though their RMSE was smallest that year. The very low 1984

TABLE 4-12

F10.7 RMSE and RMSE/SD Values for

First Day of Predictions

Year	RMSE(fcs)	RMSE(per)	RMSE/SD(fcs)	RMSE/SD(per)
1971	6.3830	4.7072	.3107	.2291
1972	5.4259	5.3943	.2554	.2539
1973	3.2461	3.6143	.2444	.2721
1974	3.6815	4.4482	.2858	.3453
1975	2.3449	2.2604	.2851	.2748
1976	2.0177	1.6759	.4328	.3595
1977	2.8114	2.9438	.2516	.2634
1978	6.1957	7.6982	.2358	.2930
1979	8.9778	10.9769	.2670	.3264
1980	9.3086	10.5688	.2699	.3064
1981	9.1393	9.9893	.2469	.2698
1982	9.2639	11.0590	.2373	.2833
1983	5.4426	5.8211	.2556	.2733
1984	3.3974	4.3936	.1384	.1790
TOTAL	6.2933	7.0774	.1212	.1363

ratio is more difficult to explain; perhaps it is due to the small sample.

The total ratio scores appear unusually low at first glance; however, the reasoning for this is easily explained. As noted above, the annual standard deviations varied almost an order of magnitude. When combined all together, the observed F10.7 daily values exhibit quite a range. Therefore, the total standard deviation, which heavily weights large departures from normal, would be expected to be large. The total standard deviation of 51.9 does not seem so unusual any more, and since that is the denominator of the ratio score, a very small value for the total ratio results.

The difference between RMSE scores increases to 1.8785 for the second day predictions and to 2.8209 for the third day (see Tables 4-13 and 4-14). The annual persistence RMSE score is only less than the forecaster's score in 1971 and 1976 for the second day and never smaller on the third prediction day. While all scores tend to increase for the longer predictions within each year, the increase is much more drastic for the years just after solar maximum (1980-1982) compared to the years around solar minimum (1975-1977). Even though this may seem significant, a comparison of the ratio values for these years would show that all values tended to increase about the same amount; in fact, all ratio values approximately doubled from the first to the third day.

Significant Error Comparison. The F10.7 significant error tables (4-15 to 4-17) provide information which is

Table 4-13

F10.7 RMSE and RMSE/SD for

Second Day of Predictions

Year	RMSE(fcs)	RMSE(per)	RMSE/SD(fcs)	RMSE/SD(per)
1971	7.8247	7.7876	.3809	.3835
1972	8.1005	9.2309	.3813	.4345
1973	5.0831	6.2008	.3827	.4669
1974	5.8773	7.2096	.4562	.5596
1975	3.3650	3.5739	.4091	.4345
1976	2.6073	2.4506	.5592	.5256
1977	3.9935	4.3642	.3573	.3905
1978	9.5431	11.9336	.3633	.4543
1979	13.0112	16.2113	.3869	.4821
1980	14.2520	17.1381	.2699	.3064
1981	14.6706	16.6982	.3963	.4511
1982	14.8900	18.9916	.3814	.4864
1983	8.0536	8.7352	.3782	.4102
1984	5.6767	7.3609	.2313	.2999
TOTAL	9.5635	11.4420	.1842	.2204

Table 4-14

F10.7 RMSE and RMSE/SD Values for

Third Day of Predictions

Year	RMSE(fcs)	RMSE(per)	RMSE/SD(fcs)	RMSE(per)
1971	10.2279	10.7214	.4979	.5219
1972	10.8252	12.7925	.5095	.6021
1973	6.7510	8.5565	.5083	.6443
1974	7.3970	9.7467	.5742	.7566
1975	4.2905	4.8709	.5217	.5922
1976	3.1735	3.1821	.6807	.6825
1977	5.0607	5.7082	.4528	.5108
1978	12.3680	16.1553	.4708	.6150
1979	16.5865	21.6012	.4932	.6424
1980	19.1189	23.3844	.5542	.6779
1981	20.2739	22.8758	.5477	.6180
1982	20.1814	26.1185	.5169	.6690
1983	10.5028	11.5078	.4932	.5404
1984	8.1503	10.1602	.3321	.4140
TOTAL	12.7458	15.5667	.2455	.2998

very easy to interpret. They will therefore be discussed as a group. First, the total forecaster errors are less than persistence in all three cases (total, under- and overforecasts) for all three days. Second, the difference increases slightly each forecast day. Third, there is a very obvious relation of significant errors to the solar cycle.

This solar cycle effect is most evident between the years of solar minimum, 1975 through 1977. During this period, both forecast methods produced a very small number of bad forecasts, essentially none in 1976 among all three days. The increase is quite dramatic approaching solar maximum. During the years 1978 through 1982, when F10.7 was fluctuating wildly (see standard deviations, Table 4-1), both methods made a significant error on the average of once every four days for one day predictions, to more than one error every other day for the third day prediction.

Although for F10.7 predictions forecasters tended to do better in both the number of under- and overforecasts, their performance on underforecasting was rarely better than three percentage points. The years 1979 and 1982 show virtually no difference between the methods. Much more skill is indicated by the forecasters' overforecast performance. The scores are lower than persistence by at least five points during the solar maximum years. As with the Ap significant error scores, it appears that forecasters do a better job predicting the end of an active event rather than the start. The results of the sign test should further demonstrate this fact. It should be noted, as with Ap, the prediction of when

Table 4-15

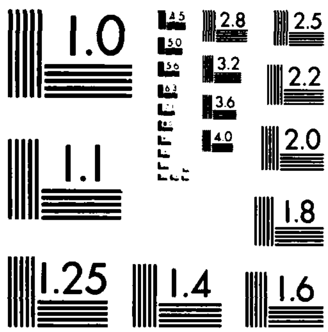
F10.7 Significant Errors (in percent) for

First Day of Predictions

Year	Total		Underforecast		Overforecast	
	Fcst	Pers	Fcst	Pers	Fcst	Pers
1971	6	4	3	2	3	2
1972	5	5	3	3	2	2
1973	1	1	1	1	0	0
1974	3	4	2	2	1	2
1975	1	0	1	0	0	0
1976	1	0	0	0	1	0
1977	1	0	0	0	1	0
1978	9	14	5	8	4	6
1979	18	25	10	12	8	13
1980	22	29	12	15	10	14
1981	23	28	13	13	10	15
1982	23	31	12	15	11	16
1983	7	6	3	3	4	3
1984	7	16	5	9	2	7
TOTAL	9	12	5	6	4	6

AD-A152 959 FORECAST VERIFICATION OF THE 107 CENTIMETER SOLAR FLUX 2/2
AND THE AP DAILY G..(U) AIR FORCE INST OF TECH
WRIGHT-PATTERSON AFB OH SCHOOL OF ENGI.. P M NOSTRAND
UNCLASSIFIED DEC 84 AFIT/GSO/PH-OS/84D-2 F/G 4/1 NL

END
FILMED
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Table 4-16

F10.7 Significant Errors (in percent) for

Second Day of Predictions

Year	Total		Underforecast		Overforecast	
	Fcst	Pers	Fcst	Pers	Fcst	Pers
1971	16	19	7	10	9	9
1972	18	25	9	12	9	13
1973	6	9	3	4	3	5
1974	9	13	7	7	2	6
1975	2	2	2	1	0	1
1976	0	0	0	0	0	0
1977	2	3	1	2	1	1
1978	21	34	11	19	10	15
1979	36	41	20	19	16	22
1980	43	56	22	28	21	28
1981	46	52	25	25	21	27
1982	44	53	23	25	21	28
1983	16	21	7	9	9	12
1984	26	41	13	24	13	17
TOTAL	21	26	11	13	10	13

Table 4-17

F10.7 Significant Errors (in percent) for

Third Day of Predictions

Year	Total		Underforecast		Overforecast	
	Fcst	Pers	Fcst	Pers	Fcst	Pers
1971	27	31	12	15	15	16
1972	25	42	11	22	14	20
1973	12	22	4	9	8	13
1974	12	23	8	12	4	11
1975	4	6	3	3	1	1
1976	0	1	0	1	0	0
1977	6	9	3	4	3	5
1978	38	51	19	28	19	23
1979	45	53	24	26	21	27
1980	54	66	28	32	26	34
1981	56	68	30	36	26	32
1982	58	65	31	30	27	35
1983	30	31	15	14	15	17
1984	44	58	22	31	22	27
TOTAL	29	36	15	18	14	18

an active region is going to appear is much more difficult- than predicting when the region will rotate to a position where the EUV will not significantly heat the atmosphere.

Significant Error Sign Test Results. Tables 4-18 to 4-20 list the number of days the paired forecasts fell into one of the four possible categories. While the null hypothesis is strongly rejected for all days using the total data set, it is interesting that the number of occassions when both forecasts were bad is consistently more than Nf. This indicates that while the forecaster does beat persistence, there is room for improvement since both forecasts are frequently poor.

Analysis of the data between the years continues to show the ability of both forecasts to perform with very few errors during the years around the pronounced solar minimum (1974 through 1977). The years associated with solar maximum (1978 through 1982) are where the number of forecaster errors are significantly less than the number of persistence errors. During the second and third prediction days, one feature stands out with respect to this period: Nf and Np remain almost constant while the days when both are bad increases almost 50 percent. The explanation for this, of course, involves the drastically increased standard deviation of the observed values. It appears that forecasters have skill to beat persistence but not enough skill to beat the fluctuating F10.7.

Table 4-18

F10.7 Significant Error Breakdown and Sign Test for

First Day of Predictions

Year	Good	Bad	Nf	Np	P-value
1971	331	6	8	17	.054
1972	340	9	8	9	.500
1973	361	2	1	1	.500
1974	349	6	8	2	.055
1975	363	1	0	1	.500
1976	364	0	0	2	.500
1977	363	1	0	1	.500
1978	309	27	24	5	.001
1979	262	50	38	15	.002
1980	233	56	51	26	.003
1981	234	52	50	29	.012
1982	222	53	61	29	.001
1983	331	15	8	11	.324
1984	73	14	15	18	.314
TOTAL	4162	287	268	148	.000

Table 4-19

F10.7 Significant Error Breakdown and Sign Test for

Second Day of Predictions

Year	Good	Bad	Nf	Np	P-value
1971	270	33	35	24	.097
1972	254	43	40	21	.001
1973	322	13	23	7	.003
1974	308	23	24	10	.013
1975	356	4	3	2	.500
1976	365	0	0	1	.500
1977	351	6	6	2	.145
1978	228	63	60	14	.000
1979	172	91	59	43	.069
1980	120	118	87	41	.000
1981	113	106	83	63	.058
1982	120	109	83	53	.007
1983	258	30	47	30	.034
1984	63	23	26	8	.002
TOTAL	3300	662	584	319	.000

Table 4-20

F10.7 Significant Error Breakdown and Sign Test for

Third Day of Predictions

Year	Good	Bad	Nf	Np	P-value
1971	215	62	48	37	.139
1972	185	68	86	27	.001
1973	271	31	50	13	.000
1974	274	39	44	8	.000
1975	340	9	12	4	.039
1976	362	0	4	0	.063
1977	324	13	20	8	.018
1978	153	111	73	28	.000
1979	118	107	85	55	.007
1980	80	155	89	42	.000
1981	63	151	96	55	.001
1982	70	155	82	58	.026
1983	199	53	59	54	.354
1984	40	42	27	11	.008
TOTAL	2694	996	775	400	.000

Absolute Error Sign Test Results. The results of the paired sign test using the absolute error scores are in Table 4-21. Once again, the forecasters statistically proved their skill over persistence for all three days. There is only one observation to make here which is somewhat unusual. Inspection of the p-values reveals that the same four years cannot accept the alternate hypothesis at the chosen level of significance: 1971, 1975, 1976 and 1983. The two middle years are during solar minimum and as already indicated, persistence has a tendency to perform well during this period. The explanation for the two outside years is less certain but it is likely to be related to the solar cycle also. The observed annual statistics for 1983 (table 4-1) reveal that a sharp drop in the mean and standard deviations occurred in that year. This drop, coupled with the fact that the 27 day solar activity cycle is most stable and predictable in the middle of the declining phase of the 11 year cycle, may be the reason for the strong showing by persistence in these years.

Table 4-21

F10.7 Absolute Value Sign Test

Year	Day 1			Day 2			Day 3		
	Nf	Np	P	Nf	Np	P	Nf	Np	P
1971	145	161	.196	171	147	.099	177	150	.075
1972	186	133	.002	212	120	.000	230	114	.000
1973	170	117	.001	206	106	.000	219	112	.000
1974	173	110	.000	187	131	.001	202	128	.000
1975	113	92	.081	132	122	.276	140	119	.107
1976	78	98	.071	120	137	.159	126	148	.103
1977	118	89	.026	158	112	.003	163	119	.005
1978	197	116	.000	225	114	.000	225	125	.000
1979	179	142	.023	187	150	.025	199	136	.001
1980	204	138	.000	224	125	.000	235	116	.000
1981	192	146	.007	198	147	.004	202	156	.009
1982	202	134	.000	216	135	.000	226	135	.000
1983	152	139	.241	160	161	.500	163	166	.456
1984	74	33	.000	81	32	.000	77	38	.000
TOTAL	2183	1648	.000	2477	1739	.000	2584	1762	.000

V. Conclusion and Recommendations

This thesis set out to do two things: to check the accuracy of AFGWC forecasts and persistence as a forecast and to conduct a test which would determine if the forecasters exhibited skill when compared to persistence as an unskilled forecast. With respect to these objectives, this research was able to provide answers. Additionally, in support of the Air Force Space Command Statement of Work (Dept of the Air Force, 1984), this report reviewed the current state of solar forecasting methods and identified future requirements and prospects for improved solar flux and geomagnetic index forecasting. This chapter will summarize the results of the analysis and conclude with a few observations and recommendations about the future of space environment forecasting.

The last chapter presented a barrage of numbers for and against the quality of the forecaster's predictions. Most of the results proved very favorable for the forecaster with the exception of the first day predictions of Ap. For all five other cases (two and three day Ap forecasts and one, two and three day F10.7 forecasts), the total results were unambiguous: as evidenced by the difference in RMSE values, the forecasters are more accurate than persistence and they most definitely make fewer significant errors and more minimum absolute errors compared with the unskilled persistence "forecasting" technique. The first day Ap forecast should not be overemphasized. The only real strike against this

forecast is its complete failure in the paired sign test of absolute error differences when the number of smaller persistence errors exceeded the forecast errors by such a margin that Np was almost declared significantly larger than Nf. However, the RMSE comparison had smaller forecaster values and the significant error sign test was able to reject the null hypothesis of no difference in favor of the SESS forecast.

It should be emphasized that in all cases the difference between forecasts increased dramatically the further out the forecast went. In other words, persistence performed worse by far on the three day forecasts. Additionally, when the solar cycle was taken into consideration by analyzing the data in annual blocks, persistence was able to show some credibility during the years when the observed standard deviation was small. However, this credibility was mainly in the failure to reject the null hypothesis, primarily on the first forecast day.

The preceeding synopsis of the results provides the basis for the author's recommendation to NORAD to begin using the forecasts produced by the space environmental forecasters at AFGWC.

This endorsement is not meant to imply that there is no room for improvement in the SESS forecasts, or in fact to suggest that better heating parameters may not become available in the future. One of the areas where the forecasters performed worst was in the prediction of sudden upswings in the Ap value. The literature acknowledged that this is a

problem (Secan and Thompson, 1979; Patterson, 1984; Joselyn, 1982). Unfortunately, the prospect for creating an ability to forecast storm and substorm commencement requires an even better understanding of the mechanisms which cause storms and a source of observations which would help to establish a basis for making storm predictions more quantitative and less qualitative. Research into this question continues (Knecht, 1984; Allen, 1984). There is general agreement that to best improve the situation a satellite would have to be placed out in the solar wind. An alternative would be to switch to another geomagnetic index which better measured disturbed geomagnetic conditions in the auroral zone such as the AE index. Unfortunately, a real time AE index is not available, and if it were, it would require the conversion of the existing atmospheric density models to accept this new index.

A similar argument is made about the F10.7 measurement's use as an index of EUV heating. A better situation would include direct measurements of the EUV from a spacecraft rather than using a parameter which does not have an excellent correlation with the EUV. Unfortunately, the prospect of getting either space system does not appear imminent.

One final comment is appropriate about the verification process currently in existence for space environment forecasts. The author agrees with Smith (1979:431) that room for improvement exists in this aspect of space forecasting. Statistical analysis techniques should be used and clearly

explained in the feedback to the user and the forecaster. The author recommends that the sign test become one such verification technique.

Appendix: F10.7 Regression Equations

This appendix contains the original and revised F10.7 regression equations for 1, 2 and 3 day predictions. The numbers in parentheses represent the day of an observation (0 or negative numbers) or a forecast (positive numbers) of F10.7. For example, F(-1) is yesterday's observation while F(0) is today's. The original equations were developed in 1966 and are identified as FO(+), FO(1) is tomorrow's forecast value. The revised equations are currently in use at AFGWC and are identified as FR(+) (Prochaska, 1984:24,26).

$$FO(1)=0.7687+1.0929*F(0)-.0454*F(-1)-.0951*F(-3)-.0375*F(-4) \\ -.0211*F(-13)+.0566*F(-15)+.0015*F(-19)+.0429*F(-23)$$

$$FO(2)=1.6063+1.1315*F(0)-.1432*F(-2)-.1173*F(-3)-.0449*F(-4) \\ -.0449*F(-13)+.1162*F(-14)+.0224*F(-19)+.0793*F(-23)$$

$$FO(3)=2.5208+1.2188*F(0)-.1516*F(-1)-.1442*F(-2)-.1924*F(-3) \\ -.0399*F(-11)+.1426*F(-14)+.0224*F(-19)+.1271*F(-22)$$

$$FR(1)=0.5461+1.0623*F(0)-.1474*F(-3)+.0217*F(-15) \\ +.0345*F(-19)+.0247*F(-26)+0.5$$

$$FR(2)=1.1426+1.0970*F(0)-.2737*F(-3)+.0670*F(-14) \\ +.0666*F(-19)+.0484*F(-25)+0.5$$

$$FR(3)=2.0766+1.1620*F(0)-.2014*F(-2)-.2009*F(-3) \\ +.0808*F(-14)+.1436*F(-19)+0.5$$

Bibliography

Abraham, Bovas and Johannes Ledolter. Statistical Methods for Forecasting. New York: John Wiley & Sons, 1983.

Akasofu, S.-I. "Development of a Geomagnetic Storm Prediction Scheme," unpublished paper presented at the Ionospheric Effects Symposium. Alexandria VA, 1-3 May 1984.

Allen, Joe H. Telephone Interview. Chief, Solar-Terrestrial Physics Division, World Data Center-A for Geomagnetism, Boulder CO, November 1984.

-----. "Some Commonly Used Magnetic Activity Indices: Their Derivation, Meaning and Use," Proceedings of a Workshop on Satellite Drag: 114-134. Boulder CO: Dept of Commerce, March 1982.

Allen, Joe H. and J. Feynman. "Review of Selected Geomagnetic Activity Indices," Solar-Terrestrial Predictions Proceedings, Volume 2, edited by R. F. Donnelly: 385-398. Boulder CO: Dept of Commerce, December 1979.

Ashton, Maj Caleb. Telephone Communication. Officer in Charge. Boulder CO: AFGWC/OL-B, May-November, 1984.

Boehm, Maj Albert R. Telephone Communication. Scientist, Tropospheric Structures Branch. Hanscom AFB MA: AFGL/LYT, October-November, 1984.

Brier, Glenn W. and Roger A. Allen. "Verification of Weather Forecasts," Compendium of Meteorology, edited by T. F. Malone. Baltimore: Waverly Press, 1951.

Campbell, W. H., et al. "Working Group Report on Geomagnetic Applications," Solar-Terrestrial Predictions Proceedings, Volume 2, edited by R. F. Donnelly: 133-136. Boulder CO: Dept of Commerce, December 1979.

Chernosky, Edwin J. et al. "The Geomagnetic Field," Handbook of Geophysics and Space Environments, edited by S. L. Valley. Air Force Cambridge Research Laboratories, Office of Aerospace Research, 1965.

Conover, W. J. Practical Nonparametric Statistics, (Second Edition). New York: John Wiley & Sons, 1980.

Dandekar, B. S. The Statistical Relations Among Q, Kp, and the Global Weather Central -K Indices. AFGL-TR-82-0010. Hanscomb AFB, MA: Air Force Geophysics Laboratories, 13 Jan 1982.

Davies, Kenneth. "Review of Recent Progress in Ionosphieric Predictions," Radio Science, 16: 1407-1430 (November-December (1981).

Department of the Air Force. Statement of Work for Solar Forecasting Reliability Analysis. SPACECOM/DO6-84-2. Peterson AFB CO: HQ AFSPACECOM, 1 March 1984.

-----. Management Analysis: Product Evaluation. AWSR 178-1. Scott AFB, IL: HQ Air Weather Service, 15 December 1983.

-----. Estimation of Daily Ap Values From 3-Hourly K Values at Selected North American Stations. Report 4586. Washington: USAF Climatic Center, Air Weather Service, May 1963.

Donnelly, Richard F. "Preface" and "Overview," Solar-Terrestrial Predictions Proceedings, Volume 2: v-viii. Boulder CO: Dept of Commerce, December 1979a.

Donnelly, Richard F., editor. Solar-Terrestrial Predictions Proceedings, Volume 1: Prediction Group Reports; Volume 2: Working Group Reports and Reviews; Volume 3: Solar Activity Predictions; Volume 4: Prediction of Terrestrial Effects of Solar Activity. Boulder CO: Dept of Commerce, December 1979b.

Dye, Capt Anita F. Personal Communication. Staff Meteorologist. Scott AFB IL: Environmental Technical Applications Center, Air Weather Service, September-October 1984.

Eis, Maj Kenneth E. Personal Communications. Chief of Operations, Office of Administration Branch, Special Support Division. Offutt AFB NE: AFGWC/WSO, July-November 1984.

-----. "Operational Problems of Satellite Drag," Proceedings of a Workshop on Satellite Drag: 5-12. Boulder CO: Dept of Commerce, March 1982.

Fraser-Smith, A. C. "Spectrum of the Geomagnetic Activity Index Ap," Journal of Geophysical Research, 77: 4209-4220 (August 1972).

Freund, John E. Modern Elementary Statistics (Fifth Edition). Englewood Cliffs NJ: Prentice-Hall, 1979.

Gringorten, Irving I., and Maj Albert R. Bohem. The B-G System of Evaluating Forecasts. AFGL-TR-82-0006. Hanscom AFB MA: Air Force Geophysics Laboratory, 4 January 1982.

Gringorten, Irving I., *et al*. "Evaluation of Forecasts with Scores Based on Climatic Probabilities," *The Collection of Papers Presented at the WMO Symposium on Probabilistic and Statistical Methods in Weather Forecasting*. 189-193. World Meteorological Organization. Symposium held in Nice, France, 8-12 September 1980.

Gringorten, Irving, I. "The Verification and Scoring of Weather Forecasts," *Journal of the American Statistical Association*, *46*: 279-296 (September 1951).

Gorney, D. J. and P. F. Mizera. "Toward an Auroral Energy Index Based on Quasi-Global X-Ray Observations." Space Division, Air Force Systems Division. Contract FO4701-82-C-00083 with Aerospace Corp. Los Angeles AFS CA, 20 July 1983.

Heckman, Gary R. "Predictions of the Space Environment Services Center," *Solar-Terrestrial Predictions Proceedings, Volume 1*, edited by R. F. Donnelly: 322-349. Boulder CO: Dept of Commerce, December 1979.

Jacchia, Luigi G. "The Earth's Upper Atmosphere -I, -II, -III," *Sky and Telescope*: 155-159, 229-232, 294-299 (March, April, May 1979).

Joselyn, Jo Ann C. "SESC Geomagnetic Predictions," *Proceedings of a Workshop on Satellite Drag*: 135-143. Boulder CO: Dept of Commerce, March 1982a.

-----."Preface" and "Discussion and Recommendations," *Proceedings of a Workshop on Satellite Drag*: v and 247-248. Boulder CO: Dept of Commerce, March 1982b.

Kraus, George. Personal Communication. Meteorological Technician, Space Environmental Support Service. Offutt AFB NE: AFGWC/WSE, April-November 1984.

Knecht, David J. Telephone Interview. Physicist, Spacecraft Interactions Branch. Hanscomb AFB: AFGL (November 1984).

Lean, J. L. "Session II Summary: Atmospheric Density Models," *Proceedings of a Workshop on Satellite Drag*: 49-50. Boulder CO: Dept of Commerce, March 1982.

Moe, Kenneth and Nancy U. Crooker. *Prediction of Geomagnetic Activity*. NASA-CR-61306. Marshall Space Flight Center, AL, August 1969 (N69-39662).

Murphy, Allan H. and Edward S. Epstein. "Verification of Probabilistic Predictions: A Brief Review," *Journal of Applied Meteorology*, *6*: 748-755 (October 1967)a.

-----. "A Note on Probability Forecasts and 'Hedging'," Journal of Applied Meteorology, 6: 1002-1003 (December 1967)b.

Nie, Norman H. et al. Statistical Package for the Social Sciences (Second Edition). New York: McGraw-Hill, 1975.

Patterson, Lt Col Verne. Telephone Interview. Chief of Aerospace Physics and Space Sciences Division, Directorate of Aerospace Development. Scott AFB IL: HQ AWS/DNXP, October-November 1984.

Paulikas, George A. and Louis J. Lanzerotti. "Impact of Geospace on Terrestrial Technology," Astronautics and Aeronautics: 42-47+ (July/August 1982).

Prochaska, Maj Robert D. Regression Versus Forecaster in Predicting Solar 10.7 Centimeter Radioflux. Student Report 84-2090. Air Command and Staff College (AU), Maxwell AFB AL, March 1984.

Prochaska, Capt Robert D., Maj Ray E. Townsend, Capt Richard W. Cannata. Source Book of the Solar-Geophysical Environment (Second Edition). Offutt AFB NE: Space Environmental Support Branch, Air Force Global Weather Central, August 1981.

Prochaska, Capt Robert D. Geomagnetic Index Calculation and Use at AFGWC. AFGWC-TN-80-002. Offutt AFB NE: Air Force Global Weather Central, April 1980.

Richmond, Arthur D. "Session III Summary: Model Inputs and Indexes, Existing and Proposed," Proceedings of a Workshop on Satellite Drag: 95-97. Boulder CO: March 1982.

Roehrich, Ronald. Telephone Communication. Director of Astrodynamics. Peterson Field AFB CO: HQ AFSPACECMD/DOA, March-November 1984.

Rostoker, Gordon. "Geomagnetic Indices," Reviews of Geophysics and Space Physics, 10: 935-950 (November 1972).

Schleher, Maj Jeffrey S. Personal Communication. Assistant Chief of Aerospace Physics and Space Sciences Division, Directorate of Aerospace Development. Scott AFB IL: HQ AWS/DNXP, May-November 1984.

Silvert, William. "The Optimization and Evaluation of Forecasts," The Collection of Papers Presented at the WMO Symposium on Probabilistic and Statistical Methods in Weather Forecasting. World Meteorological Organization Symposium held in Nice, France, 8-12 September 1980.

Smith, J. B. "Summary of Forecasters' Meeting on April 25, 1979, During the International Solar-Terrestrial Predictions Workshop," _Solar-Terrestrial Predictions Proceedings, Volume 1_, edited be R. F. Donnelly: 428-432. Boulder CO: Dept of Commerce, December 1979.

Springer, Capt Bruce D. "Sunspots and 10.7-cm Solar Microwave Flux: Predictive Techniques," _Proceedings of a Workshop on Satellite Drag_: 99-113. Boulder CO: Dept of Commerce, March 1982.

Thompson, Lt Col R. L. and Capt J. A. Secan. "Geophysical Forecasting at AFGWC," _Solar-Terrestrial Predictions Proceedings, Volume 1_, edited by R. F. Donnelly: 350-366. Boulder CO: Dept of Commerce, December 1979.

Townsend, Maj Ray E. Telephone communication. Chief of Advanced Development Branch. Sunnyvale AFS CA: Det 3, HQ AWS, April-November 1984.

Vampola, A. L., _et al_. "Working Group Report on: User Requirements for Solar-Terrestrial Predictions for Spacecraft Applications," _Solar-Terrestrial Predictions Proceedings, Volume 2_, edited by R. F. Donnelly: 1-20. Boulder CO: Dept of Commerce, December 1979.

Vita

Philip Michael Nostrand entered this world on 13 July, 1959. He was raised in East Northport, New York and graduated Northport High School in 1977. He attended Cook College, Rutgers University in New Brunswick, New Jersey on a four year ROTC scholarship. Philip was commissioned and graduated in 1981 with a Bachelor of Science Degree in Meteorology. Sunnyvale AFS was his first assignment where he served as Assistant Chief, Operations Branch of Det 3, HQ AWS. His duties included launch, on-orbit and recovery support for Air Force satellite systems. One aspect of this job included using AFGWC Ap and F10.7 forecasts to predict satellite drag on low orbiting satellites. After 18 months in sunny California he was selected to attend the Air Force Institute of Technology School of Engineering to study in the Space Operations master's program in May 1983.

Permanent address: 907 5th St
East Northport, NY 11731

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION/DOWNGRADING SCHEDULE

3. DISTRIBUTION/AVAILABILITY OF REPORT
Approved for public release;
distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
AFIT/GSO/PH-OS/84D-2

5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION
School of Engineering

6b. OFFICE SYMBOL *(If applicable)*

7a. NAME OF MONITORING ORGANIZATION

6c. ADDRESS *(City, State and ZIP Code)*
Air Force Institute of Technology
Wright-Patterson AFB, Ohio 45433

7b. ADDRESS *(City, State and ZIP Code)*

8a. NAME OF FUNDING/SPONSORING ORGANIZATION

8b. OFFICE SYMBOL *(If applicable)*

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER

8c. ADDRESS *(City, State and ZIP Code)*

10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.

11. TITLE *(Include Security Classification)*
See Box 19

12. PERSONAL AUTHOR(S)
Philip M. Nostrand, B.S., 1st Lt, USAF

13a. TYPE OF REPORT
MS Thesis

13b. TIME COVERED
FROM ________ TO ________

14. DATE OF REPORT *(Yr., Mo., Day)*
1984 December

15. PAGE COUNT
113

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB. GR.
04	01	

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Forecasting, Geomagnetism, Upper atmosphere
Solar radiation, Low orbit trajectories

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

Title: FORECAST VERIFICATION OF THE 10.7 CENTIMETER SOLAR FLUX
AND THE Ap DAILY GEOMAGNETIC ACTIVITY INDICES

Thesis Chairmen: Joseph W. Coleman, Lieutenant Colonel, USAF
James J. Lange, Major, USAF

Approved for public release: IAW AFR 190-17.
21 Feb 85
LYNN E. WOLAVER
Dean for Research and Professional Development
Air Force Institute of Technology (ATC)
Wright-Patterson AFB OH 45433

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT
UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS ☐

21. ABSTRACT SECURITY CLASSIFICATION
UNCLASSIFIED

22a. NAME OF RESPONSIBLE INDIVIDUAL
James J. Lange, Major, USAF

22b. TELEPHONE NUMBER *(Include Area Code)*
513-255-4498

22c. OFFICE SYMBOL
AFIT/ENP

Air Force Global Weather Central space environmental forecasts of the 10.7 centimeter solar radioflux index and the daily average geomagnetic activity index were compared with persistence "forecasts" to check for accuracy and skill. One, two and three day forecasts were compared. The AFGWC forecasts were found to be more accurate and skillful than the persistence forecasts.

The data base covered the period from 4 January 1971 through 29 April 1984. Statistics were calculated for the total data set and each individual year. Root mean square error and percentage of significant errors were used as measures of accuracy. A paired sign test was used to compare for skill. The test was run on significant errors and absolute errors. A significant error is when the difference between the forecast value and the verifying observed value (ie. the observation one, two or three days hence) is greater than ten.

The total data base yielded results which favored the AFGWC forecasts in all instances except one. The exception was the one day Ap sign test on absolute errors. Persistence also tended to do as well or better than the AFGWC forecasts on some individual years, primarily during the years around solar minimum (1975-1977) and also for the one day forecasts. It was found that AFGWC performed better at predicting a sudden decrease in the index values than it did predicting a sudden increase.

END

FILMED

5-85

DTIC